



## 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

이 학 박 사 학 위 논 문

Self-Consistent Estimator of Marginal Distribution  
and Two-Sample Problems  
for Interval-Valued Data

구간자료에 대한 자기일치 분포 추정량과  
이표본 문제들

2019년 8월

서울대학교 대학원

통계학과

최 혜 정

**Self-Consistent Estimator of Marginal Distribution  
and Two-Sample Problems  
for Interval-Valued Data**

**By**

**Hyejeong Choi**

**A Thesis**

**submitted in fulfillment of the requirement  
for the degree of  
Doctor of Philosophy  
in Statistics**

**Department of Statistics  
College of Natural Sciences  
Seoul National University  
August, 2019**

## ABSTRACT

# Self-Consistent Estimator of Marginal Distribution and Two-Sample Problems for Interval-Valued Data

Hyejeong Choi

The Department of Statistics

The Graduate School

Seoul National University

This thesis is composed of three subjects on the analysis of interval-valued data. First, we propose a new type of marginal distribution estimator, named as a self-consistent estimator (SCE) and investigate its properties. Second, we propose several new approaches to compare two interval-valued samples, and also propose a procedure to test the stochastic order between two samples.

In interval-valued data, the variable of interest is provided in the form of a two-dimensional vector of lower and upper bounds, not a single value. It is of interest to represent interval-valued data with a univariate random variable/marginal distribution. Two estimators, the empirical histogram estimator, and nonparametric kernel estimator have been proposed for the estimation of the marginal histogram in the literature. In the first part of the thesis, we define a new marginal representation, named as self-consistent marginal, for interval-valued data, and propose an SCE to estimate it. In the second and third parts of the thesis, we discuss

how to compare two samples of interval-valued data. One is about the equality of two samples, and the other is about the stochastic order between two. First, to test equality, we consider four methods. Two are based on the existing approach for bivariate data, and the other two are newly proposed based on the univariate marginalization of interval-valued data. Second, to test the stochastic order, we propose a test statistic which belongs to U-statistic and derive its asymptotic null distribution. We conduct a comprehensive numerical study to investigate the performance of the newly proposed methods along with the existing methods. We further illustrate the advantages of the proposed methods over the existings by applying to empirical examples.

**Keywords:** Interval-valued data; marginalization; nonparametric distribution function; self-consistent estimator; two-sample test; stochastic order; blood pressure data

**Student Number:** 2015 – 30970

# Contents

<b>Abstract</b>	<b>i</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Self-Consistent Estimator of Marginal Distribution</b>	<b>6</b>
2.1. Introduction . . . . .	6
2.2. Review the existing methods . . . . .	8
2.3. Self-consistent estimator . . . . .	10
2.3.1. Estimator . . . . .	10
2.3.2. Innermost intervals . . . . .	13
2.3.3. The almost sure limit of the SCE: self-consistent marginalization . . . . .	15
2.4. Numerical study . . . . .	21
2.4.1. Continuous cases . . . . .	22
2.4.2. Discrete cases . . . . .	25
2.4.3. The sensitivity to the coarseness of the in- tervals . . . . .	28
2.5. Data examples . . . . .	33
2.5.1. Rally data . . . . .	33
2.5.2. Blood pressure data . . . . .	35

2.6.	Conclusion . . . . .	37
2.6.1.	Extension to p-dimensional interval-valued data . . . . .	38
<b>3</b>	<b>Two-Sample Tests for Interval-Valued Data</b>	<b>41</b>
3.1.	Introduction . . . . .	41
3.2.	Methods . . . . .	42
3.2.1.	Combined (CB) test . . . . .	42
3.2.2.	Hotelling's $T^2$ (HT) test . . . . .	44
3.2.3.	Marginalization-based (UK and GK) test . . . . .	45
3.3.	Numerical study . . . . .	48
3.3.1.	Normal distribution with equal covariances . . . . .	48
3.3.2.	Non-normal cases . . . . .	53
3.3.3.	Normal distribution with unequal covariances . . . . .	57
3.4.	Data example . . . . .	61
3.4.1.	Sub-sampling . . . . .	62
3.5.	Conclusion . . . . .	64
<b>4</b>	<b>Testing for Stochastic Order in Interval-Valued Data</b>	<b>66</b>
4.1.	Introduction . . . . .	66
4.2.	Simple stochastic order . . . . .	67
4.3.	Test statistic . . . . .	69
4.4.	Numerical study . . . . .	72
4.5.	Data example . . . . .	76
4.6.	Conclusion . . . . .	76
<b>5</b>	<b>Conclusion</b>	<b>78</b>
	<b>References</b>	<b>80</b>





# List of Tables

2.1	Simulation results for continous cases. The distance between the estimated distribution ( $\widehat{F}(x)$ ) and target distribution ( $F_T(x)$ with $T = "C", "A", "S"$ ) is displayed with the standard errors in parentheses. . . . .	24
2.2	$dG(I)$ , $\mathbf{p}_s$ and $\mathbf{p}_A$ for (D.DIS). . . . .	26
2.3	$dG(I)$ and the corresponding $\mathbf{p}_s$ and $\mathbf{p}_A$ for cases (D.OVL1) and (D.OVL2). . . . .	26
2.4	Simulation results for discrete case. The estimation errors with their standard errors in parentheses. . . . .	27
2.5	Data for each case. . . . .	30
2.6	The innermost intervals and SCE for each case. The numbers in parentheses([ ]) to the right of case notation indicate the total number of innermost intervals. . .	31
2.7	Innermost intervals and correponding SCE. Intervals with masses greater than 0.01 are displayed. The numbers in parentheses are the GK estimates. . . . .	34

2.8	Descriptive statistics of the BP data by race. The p-value is from the t-test on the alternative hypothesis that the BP of African-American is higher than that of Caucasian. At the first column, mid-BP indicates the center of the BP data. . . . .	36
3.1	Summary of the settings where $\Sigma = (1 \quad \rho; \rho \quad 1)$ . At the first column, the left character of the hyphen (-) denotes the distribution of $(C, \log R)$ and the right represents the difference between the two populations. Among the left, N indicates “normal”, T for “T with df 5”, and SN for “skew-normal”. Among the right, C represents “mean of center”, R for “mean of range”, C.S for “mean and skewness of center”, COV for “covariance”, C.V for “mean and variance of center”, and R.V for “mean and variance of range”. The first population is denote by $\Pi_1$ and the second is denoted by $\Pi_2$ with $\mu_1, \mu_2$ mean parameters, $\Sigma_1, \Sigma_2$ covariance matrices, and $\gamma_1, \gamma_2$ skewness parameters, respectively. . . . .	49
3.2	Simulation results. Power of each test in case of the bivariate normal distribution with equal covariances. . . .	51
3.3	Simulation results. Power of each test in case of the bivariate t-distribution with df 5 with equal covariances . . .	54
3.4	Simulation results. Power of each test in case of the bivariate skew-normal distribution with equal covariances. . . .	56
3.5	Simulation results. Power of each test in case of the bivariate normal distribution with unequal covariances. . . .	59

3.6	Summary of the results. The best and worst tests are represented for each case. At the second column, the left character of the hyphen (-) denotes the distribution of $(C, \log R)$ and the right represents the difference between the two populations. . . . .	61
3.7	Two-sample tests for the whole BP data. . . . .	62
3.8	Descriptive statistics of $(C, \log R)$ for the BP data by race, where $C$ is the center and $R$ is the half-range. This is a summary of the population of the sub-samples. The correlation coefficient is for the center and log-transformed half-range. . . . .	63
3.9	Power of four two-sample testing methods for different sub-sample sizes. . . . .	64
4.1	Simulation results for the stochastic order tests. The power of each test is displayed. At the first column, the character denotes the distribution of $(C, \log R)$ : N indicates “normal” and T indicates “ $t$ -distribution with df 5”. At the second row, U-perm and U-asym represent the stochastic order tests using the U-test, where “perm” and “asym” imply the null distribution is approximated by a permutation method and the asymptotic result in Theorem 4.2, respectively. B-KS denotes the bivariate K-S test. . . . .	75
4.2	Two-sample order tests for the BP data . . . . .	76

# List of Figures

2.1	Innermost intervals. . . . .	14
2.2	Nearly-overlapped intervals. . . . .	18
2.3	Display of the marginal distribution functions $F_C$ , $F_A$ and $F_s$ for each case. . . . .	23
2.4	Boxplots and density plots for (D.DIS). The third plot represents the estimated probability density functions. The height of the bar represents the average probability that the SCE(=UK) estimates for each interval for the 100 data sets. The standard error bar is also plotted. The red line indicates the true probability $\mathbf{p}_s(= \mathbf{p}_A)$ . The gray line is the average of the probability density function by the GK for the 100 data sets. . . . .	28
2.5	Boxplots for (D.OVL1) and (D.OVL2). . . . .	29
2.6	Rally data and the SCE and GK. In the SCE plot, the height of the bar indicates the size of the mass, and the width of the bar indicates the width of the innermost interval. . . . .	34
2.7	Density plots by RACE : Race = 1 indicates Caucasians, and RACE = 2 indicates African-Americans. . . . .	37

3.1	Contour plots of the two groups of BP data . . . . .	62
4.1	A graphical illustration of the order of interval-valued data	68
4.2	A graphical illustration of the setting of numerical study	74

# Chapter 1

## Introduction

In interval-valued data, the variable of interest is not represented as a single value but provided in the form of an interval with lower and upper bounds. We are exposed to various interval-valued data in practice. Two examples are, first, monthly stock prices of individual companies are often summarized as the peak and bottom prices during a month. Second, in a survey of demographic variables, a person's income is recorded as an interval to protect his/her privacy. Two examples in the above are very different in their generation process. The first example is the precise description of a set-valued entity, whereas the second example is an imprecise description of a point-valued quantity (Couso and Dubois, 2014; Blanco-Fernández and Winker, 2016). Blanco-Fernández and Winker (2016) name the first example type of interval-valued data as “min-max (MM)” or “ontic” data and the second example type as “measurement error (ME)” or “epistemic” interval-valued data.

The ME-type assumes there exists a true value and the true value is not observable directly, but only observable as an interval.

This situation may happen due to different reasons. Sometimes the exact value is not available due to confidentiality issues as stated above. Also, the precise value of a variable might not be obtained due to the use of non-sufficiently accurate measurement device. In this case, because of the possibility of errors in the observation of experimental data, it is more appropriate to provide the uncertainty of the observed values, which leads to interval-valued data. Interval-censored data that many clinical trials and longitudinal studies may generate is also one of this type. By interval censoring, the failure time is not observed precisely but is known to be lying in an interval obtained from a sequence of inspection times.

The MM-type is the case where the interval itself is the object of interest. This type of interval-valued data is generated when aggregating large datasets to the minimum and maximum values or focusing on the range of variation of the variable. The stock price data above-mentioned is an example of the first case, which is a kind of summary data. Summarized data is an example of symbolic data, and summarization of a dataset can also be represented by lists, histograms, and so on (Billard and Diday, 2003). Some studies focus on the range of variation of the variables over a specified period or within a cross-section, leading to the MM-type interval-valued data. Blood pressure data usually recorded in maximum and minimum during a heartbeat cycle can be the example.

We can represent interval-valued data in two coordinate systems: L-U system and C-R system. The L-U system is a general notation for interval-valued data, where interval data is defined as

an element of

$$\{(L, U] : L, U \in \mathcal{R}, L < U\}$$

for a half-open case. Intervals can also be represented as open or closed. In the L-U system, the restriction,  $L < U$  introduces difficulties in statistical modeling to data. To avoid this problem, the transformation of data is proposed. An equivalent representation of an interval is given by the center (mid-point) and radius (half-range) of the interval, namely,  $(C, R)$  where  $C = (L + U)/2$  and  $R = (U - L)/2$ . We can also replace  $R$  with  $\log R$  to remove the constraint,  $R > 0$ . We call this notation as the C-R system.

While the same notation is used for both the MM-type and ME-type interval-valued data, the analysis and inference in both types should be different (Blanco-Fernández and Winker, 2016; Grzegorzewski, 2018). In the ME-type, we deal with usual real-valued random variables, but the problem is that the realization of the value is not precise but obtained as an interval. Thus the statistical analysis is based on this imprecise information about the point data, and the result may be expressed in an imprecise way. For example, parameter estimates are often represented as a set of all possible values under the interval uncertainty. On the other hand, in the MM-type, we focus on the random interval itself, not the underlying variable. In this case, detailed modeling or probabilistic approaches are performed for the lower and upper bounds (corresponding to the center and half-range of the C-R system).

Many studies have been conducted on the ME-type data in the name of fuzzy data or interval censoring, but research on the



MM-type data is relatively insufficient. In the thesis, we focus on studying the MM-type interval-valued data and deal with the following three subjects:

1. Self-consistent estimator of marginal distribution
2. Two-sample tests for interval-valued data
3. Testing for stochastic order in interval-valued data.

The first is about the methods to find a single-valued representation of interval-valued data, and the other two subjects are about the approaches to compare two interval-valued samples, one for the equality test and the other for the order test. In this thesis, the results of research on the above subjects are arranged in order. The organization of the paper and brief contents of each chapter are as follows.

In Chapter 2, we define a new marginal representation, named as a self-consistent marginal, for interval-valued data, and propose a self-consistent estimator (SCE) to estimate it. It is of interest to find a marginal (single-valued) representation  $X$  for interval-valued data  $(L, U]$ , which is composed of two random vectors. More specifically, the estimation of the distribution function of  $X$  expressed as  $F(x)$  is addressed. We refer to this marginal representation as marginalization. We theoretically and numerically investigate the properties of the SCE under various assumptions. We further illustrate the advantages of the SCE over the two marginal histogram estimators with empirical examples. In Chapter 3, methods to compare two samples of interval-valued data are discussed. We consider four methods, two of which are the appli-

cations of existings for bivariate data and the other two are based on the univariate marginalization of interval-valued data. We conduct a comprehensive numerical study and analysis of real data to understand the performance of four methods. In Chapter 4, we construct a procedure to test the stochastic order of two samples of interval-valued data. We propose a test statistic which belongs to U-statistic and derive its asymptotic distribution under the null hypothesis. We compare the performance of the newly proposed method with the existing one-sided bivariate K-S test using real data and simulated data. Finally, we conclude the thesis in Chapter 5.

## Chapter 2

# Self-Consistent Estimator of Marginal Distribution

### 2.1. Introduction

As we mentioned at the beginning, we aim to find a marginal representation  $X$  for interval-valued data  $(L, U]$ , which is intrinsically a two-dimensional random vector. More specifically, researchers are interested in the estimation of the distribution function of  $X$ , denoted by  $F(x)$ . For the ME-type interval data,  $X$  is the true value with no ME. Under the independence between  $X$  and  $(L, U]$ , the nonparametric maximum likelihood estimator (MLE) of  $F(x)$  is studied much with the name of interval censoring (Gentleman and Geyer, 1994; Gómez et al., 2004; Lim et al., 2009; Peto, 1973; Turnbull, 1976; Wong and Yu, 1999; Yu et al., 2000). For the MM-type data,  $X$  is not uniquely and intuitively defined, and several dif-

ferent versions of marginalization are introduced, which include the center of the interval as  $X_C = (L + U)/2$  (Billard and Diday, 2000) and the marginally histogrammed variable  $X_m$  (Bertrand and Goupil, 2000), which will be explained in more detail in Section 2.2.

In this chapter, we consider the self-consistent estimator (SCE) of interval-valued data and define the self-consistent marginal as its almost sure limit. Suppose we observe  $n$  independent intervals  $\{I_i = (\ell_i, u_i], i = 1, \dots, n\}$ . The SCE is defined as the solution to the following equation: for any real interval  $(a, b]$

$$\int_{(a,b]} f(t) dt = \frac{1}{n} \sum_{i=1}^n \int_{(a,b]} \left\{ \frac{f(t)}{\int_{\ell_i}^{u_i} f(s) ds} \mathbf{I}(t \in (\ell_i, u_i]) \right\} dt, \quad (2.1)$$

where  $a$  and  $b$  are the elements of  $\{\{\ell_i\}_{i=1}^n, \{u_i\}_{i=1}^n\} \cup \{-\infty, \infty\}$ .

The SCE for the ME-type data is mainly well studied for interval-censored data. The SCE aims to estimate the distribution function of the non-censored variable and is known to consistently estimate it under the assumptions of non-informativeness of the interval on the non-censored true variable (Gentleman and Geyer, 1994; Yu et al., 2000). The goal of this paper is to understand the SCE estimator and its limit for the MM-type data.

The remainder of the chapter is organized as follows. In Section 2.2, we review the existing marginalization methods. In Section 2.3, we define the SCE for interval-valued data as the solution of a recursive equation. We further define a self-consistent marginalization of interval-valued data as the limit of the SCE. In Section 2.4, we numerically compare the SCE with the other two existing marginalization methods. In Section 2.5, we apply our SCE and

existing methods to the rally data and blood pressure data. In Section 2.6, we conclude the chapter with a summary.

## 2.2. Review the existing methods

In this section, we introduce two existing estimators for marginalization, the empirical histogram estimator (also known as marginal histogram estimator or the kernel estimator with the uniform kernel (UK)) and the Gaussian kernel estimator (GK) by Jeon et al. (2015).

Before introducing the two kernel estimators, we mention the key notations. Suppose we observe  $n$  independent intervals  $\{I_i = (\ell_i, u_i], i = 1, \dots, n\}$  from a population with (unknown) cumulative distribution function  $G$ . We define the basis intervals  $(\xi_k, \xi_{k+1}], k = 0, 1, \dots, K$ , where  $-\infty = \xi_0 < \xi_1 < \xi_2 < \dots < \xi_K < \xi_{K+1} = \infty$  are the unique ordered elements of  $\{\{\ell_i\}_{i=1}^n, \{u_i\}_{i=1}^n\} \cup \{-\infty, \infty\}$ .

The empirical histogram estimator is defined as follows (Bertrand and Goupil, 2000):

$$f_n^{UK}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{u_i - \ell_i} \mathbf{I}(\ell_i < x \leq u_i). \quad (2.2)$$

It assumes that the value for  $X_i$  is uniformly distributed in the interval  $(\ell_i, u_i], i = 1, \dots, n$ , and the marginalization is represented as the uniform mixture of  $n$  uniform distributions. It is also known as a marginal histogram estimator or kernel estimator with the uniform kernel. We refer to this estimator as the uniform kernel estimator (UK). The UK provides a histogram-type density with the basis intervals as bins.

Jeon et al. (2015) improve the uniform kernel estimator by imposing some structures on the distribution of data. The proposed estimator is a mixture of  $n$  univariate normal densities. That is,

$$f_n^{GK}(x) = \frac{1}{n} \sum_{k=1}^n \phi(x|\hat{\mu}_k, \hat{\sigma}_k), \quad (2.3)$$

where  $\phi(\cdot|\hat{\mu}_k, \hat{\sigma}_k)$  is the univariate normal density with mean  $\hat{\mu}_k$  and standard deviation  $\hat{\sigma}_k$  computed by

$$\begin{aligned} \hat{\mu}_k &= \frac{1}{n} \sum_{i=1}^n w_{ki} m_i, \quad \hat{\sigma}_k^2 = \frac{1}{n} \sum_{i=1}^n w_{ki} v_i, \\ m_i &= (\ell_i + u_i)/2, \quad v_i = (u_i - \ell_i)^2/12. \end{aligned}$$

The local weights  $w_{ki}$  are determined as follows. Using the centers of the intervals, we calculate Euclidean distances between  $k$ th and  $i$ th intervals, say  $d_{ki} = d_{ik}$ , and sort the distances. Let  $R_{ki}$  be the rank of the  $d_{ki}$  among  $\{d_{k1}, d_{k2}, \dots, d_{kn}\}$  with  $R_{kk} = 1$ . The weights are determined such that

$$w_{ki} \propto \frac{1}{h} \phi\left(\frac{R_{ki} - 1}{h}\right) \quad \text{and} \quad \sum_{i=1}^n w_{ki} = 1,$$

where  $\phi$  is the standard normal density and  $h$  is the bandwidth. Choosing the bandwidth  $h$  is important and Jeon et al. (2015) propose to use the Kullback-Leibler loss as follows:

$$- \int f_n^{UK}(x) \log f_n^{GK}(x) dx,$$

where  $f_n^{UK}(x)$  is the uniform kernel estimator in (2.2). Then,  $h$  is chosen to minimize the cross-validated empirical Kullback-Leibler loss

$$CV(h) = -\frac{1}{n} \sum_{i=1}^n \frac{1}{u_i - \ell_i} \int_{\ell_i}^{u_i} \log f_n^{GK(i)} dx, \quad (2.4)$$

where  $f_n^{GK(i)}(x)$  represents the Gaussian kernel estimator in (2.3) with the  $i$ -th interval left out. Jeon et al. (2015) propose to use a smoothing parameter  $\alpha$  to control the number of observations that hold nontrivial weights. For a given  $\alpha$  in the range of  $(1/n, 1]$ ,  $h$  is chosen to meet the relation  $([n\alpha] - 1)/h = 2$ , where  $[a]$  is the largest integer less than or equal to  $a$ . To improve the selection of the bandwidth, Park et al. (2016) take a scale-space approach and develop a SiZer (Significant ZERo crossing of the derivatives) tool for interval-valued data. We refer to the estimator suggested by Jeon et al. (2015) as GK named after the Gaussian kernel.

Both kernel estimators aim to estimate the *marginal aggregation of intervals* as its almost sure limit. The marginal aggregated distribution (or histogrammed marginalization) is defined as

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \frac{1}{u_i - \ell_i} \mathbf{I}(\ell_i < x \leq u_i). \quad (2.5)$$

## 2.3. Self-consistent estimator

### 2.3.1. Estimator

We propose a self-consistent method to find the marginal density of interval-valued data. The self-consistent estimate is a fixed point of the equation (2.1) in the sense that, SCE is reproduced in the righthand-side if it is plugged-in the lefthand-side of (2.1). This fixed point interpretation explains the term “self-consistent”. We construct an iterative procedure based on data and determine its exact fixed point in this section.

Suppose we observe  $n$  independent random intervals  $\{I_i = (\ell_i, u_i], i = 1, \dots, n\}$  from a population with (unknown) cumula-

tive distribution function  $G$ . The SCE, which we notate as  $\widehat{f}_s(x)$ , is defined as the solution to the estimating equation,

$$f(x) = \frac{1}{n} \sum_{i=1}^n \frac{f(x) \mathbf{I}(x \in (\ell_i, u_i])}{\int_{\ell_i}^{u_i} f(t) dt}. \quad (2.6)$$

As an iterative procedure to find  $\widehat{f}_s(x)$ , on the  $r$ -th procedure  $f_r$  is calculated by

$$f_r(x) = \frac{1}{n} \sum_{i=1}^n \frac{f_{r-1}(x) \mathbf{I}(x \in (\ell_i, u_i])}{\int_{\ell_i}^{u_i} f_{r-1}(t) dt}.$$

In terms of the conditional expectation,

$$f_r(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{E}_{r-1} [\mathbf{I}(X_i \in dx) | (\ell_i, u_i], f_{r-1}], \quad (2.7)$$

which is the self-consistent algorithm following Efron (1967).

Now we discuss the characterization of the solution to (2.6).

**Theorem 2.1.** *The solution to (2.6) is only identifiable up to the basis intervals  $(\xi_k, \xi_{k+1}]$ ,  $k = 0, 1, \dots, K$  and (2.6) is simplified to*

$$w_k = \frac{1}{n} \sum_{i=1}^n \frac{\alpha_{ik} w_k}{\sum_{l=0}^K \alpha_{il} w_l}, \quad 0 \leq k \leq K, \quad (2.8)$$

where

$$\alpha_{ik} = \mathbf{I}((\xi_k, \xi_{k+1}] \subseteq I_i) \quad \text{and} \quad w_k = \int_{\xi_k}^{\xi_{k+1}} f(x) dx.$$

*Proof.* First, we show that the solution to (2.6) is only identifiable up to the basis intervals. Suppose there is an  $\widehat{f}_1$ , which is not equal to  $\widehat{f}_s$  but has the same probabilities for the basis intervals as  $\widehat{f}_s$ . That is,

$$\int_{\xi_k}^{\xi_{k+1}} \widehat{f}_1(x) dx = \int_{\xi_k}^{\xi_{k+1}} \widehat{f}_s(x) dx, \quad \text{for } k = 0, 1, 2, \dots, K.$$



In addition, we assume that  $\widehat{f}_1(x) = 0$  for  $x \notin I_i, i = 1, \dots, n$ .

First, for any  $x \in \mathcal{R}$ , suppose  $x \in I_i$  for some  $i$ 's. Since any interval  $I_i$  is split into several consecutive basis intervals  $(\xi_k, \xi_{k+1}]$ ,  $k = s_{i_1}, \dots, s_{i_{K_i}}$ , where  $I_i = \bigcup_{k=s_{i_1}}^{s_{i_{K_i}}} (\xi_k, \xi_{k+1}]$ ,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \frac{\widehat{f}_1(x) \mathbf{I}(x \in (\ell_i, u_i])}{\int_{\ell_i}^{u_i} \widehat{f}_1(t) dt} &= \frac{1}{n} \sum_{i: x \in (\ell_i, u_i]} \frac{\widehat{f}_1(x)}{\int_{\ell_i}^{u_i} \widehat{f}_1(t) dt} \\ &= \frac{1}{n} \sum_{i: x \in (\ell_i, u_i]} \frac{\widehat{f}_1(x)}{\sum_{k=s_{i_1}}^{s_{i_{K_i}}} \int_{\xi_k}^{\xi_{k+1}} \widehat{f}_1(t) dt} \\ &= \frac{1}{n} \sum_{i: x \in (\ell_i, u_i]} \frac{\widehat{f}_1(x)}{\sum_{k=s_{i_1}}^{s_{i_{K_i}}} \int_{\xi_k}^{\xi_{k+1}} \widehat{f}_s(t) dt} \\ &= \frac{1}{n} \sum_{i: x \in (\ell_i, u_i]} \frac{\widehat{f}_1(x)}{\int_{\ell_i}^{u_i} \widehat{f}_s(t) dt} = \widehat{f}_1(x), \end{aligned}$$

where the last equality comes from  $\frac{1}{n} \sum_{i: x \in (\ell_i, u_i]} \frac{1}{\int_{\ell_i}^{u_i} \widehat{f}_s(t) dt} = 1$ . Thus,  $\widehat{f}_1(x)$  is also the solution to (2.6). Second, suppose  $x \notin I_i$ , for all  $i = 1, \dots, n$ , then  $\widehat{f}_1(x) = \widehat{f}_s(x) = 0$  satisfies the equation (2.6).

Next, we show that the equation (2.6) is equivalent to the equation (2.8). Note that both solutions to (2.6) and (2.8) assign weights to the basis intervals and only indentifiable up to the basis intervals. First, we show that the solution to (2.6) satisfies (2.8). Let  $\nu_k = \widehat{F}_s(\xi_{k+1}) - \widehat{F}_s(\xi_k)$  where  $\widehat{F}_s(x) = \int_{-\infty}^x \widehat{f}_s(t) dt$ . Then it suffices to show that  $\nu_k$ s satisfy (2.8). Since  $\widehat{F}_s(x) = \frac{1}{n} \sum_{i=1}^n \frac{\int_{\ell_i}^x \widehat{f}_s(t) dt}{\int_{\ell_i}^{u_i} \widehat{f}_s(t) dt} \mathbf{I}(t \in I_i) + \frac{1}{n} \sum_{i=1}^n \mathbf{I}(u_i \leq x)$  and  $(\xi_k, \xi_{k+1}] \subseteq I_i$  or  $(\xi_k, \xi_{k+1}] \cap I_i = \emptyset$ ,

$$\begin{aligned} \nu_k &= \frac{1}{n} \sum_{i=1}^n \frac{\int_{\xi_k}^{\xi_{k+1}} \widehat{f}_s(t) dt}{\int_{\ell_i}^{u_i} \widehat{f}_s(t) dt} \mathbf{I}(t \in I_i) \\ &= \frac{1}{n} \sum_{i: (\xi_{k+1}, \xi_k] \subseteq (\ell_i, u_i]} \frac{\int_{\xi_k}^{\xi_{k+1}} \widehat{f}_s(t) dt}{\int_{\ell_i}^{u_i} \widehat{f}_s(t) dt} = \frac{1}{n} \sum_{i=1}^n \frac{\alpha_{ik} \nu_k}{\sum_{l=0}^K \alpha_{il} \nu_l}. \end{aligned}$$

This is the same as the equation for  $w_k$ s, that is (2.8). Similarly, we can show that  $w_k$ s satisfy (2.6).  $\square$

### 2.3.2. Innermost intervals

We find that the solution to the recursive equation (2.8) is not unique. For the ME-type data, in the literature of interval-censored data, the solution that maximizes the non-parametric likelihood function of data places a mass only on a subset of (the set of) basis intervals, named as innermost intervals. The innermost intervals by Peto (1973) and Turnbull (1976) are defined as follows. Let  $\{I_i = (l_i, u_i], i = 1, \dots, n\}$  be the observed intervals. From the sets of  $\mathcal{L} = \{\ell_i, 1 \leq i \leq n\}$  and  $\mathcal{U} = \{u_i, 1 \leq i \leq n\}$ , we derive all the disjoint intervals whose left and right end-points lie in the sets  $\mathcal{L}$  and  $\mathcal{U}$  respectively, and which contain no members of  $\mathcal{L}$  or  $\mathcal{U}$  excepts at their left and right end-points, respectively. We write these intervals  $(t_1, s_1], (t_2, s_2], \dots, (t_M, s_M]$  where  $t_1 < s_1 \leq t_2 < s_2 \leq \dots \leq t_M < s_M$ ,  $M \leq n$ . Thus, for any pair of intervals  $I_i = (\ell_i, u_i]$  and  $(t_j, s_j]$ , either  $(t_j, s_j] \subseteq I_i$  or  $(t_j, s_j] \cap I_i = \emptyset$  holds.

Let us illustrate the construction of innermost intervals by an example. Suppose  $n = 6$  and the observed intervals are  $(0, 1]$ ,  $(2, 4]$ ,  $(4, 8]$ ,  $(0, 3]$ ,  $(2, 9]$  and  $(4, 6]$ . The non-trivial basis intervals are as follows:  $(0, 1]$ ,  $(1, 2]$ ,  $(2, 3]$ ,  $(3, 4]$ ,  $(4, 6]$ ,  $(6, 8]$ ,  $(8, 9]$ . Following the construction,  $\mathcal{L} = \{0, 2, 4\}$  and  $\mathcal{U} = \{1, 3, 4, 6, 8, 9\}$ . Then, there are 3 innermost intervals:  $(0, 1]$ ,  $(2, 3]$ ,  $(4, 6]$ . Figure 2.1 shows the graphical illustration of the innermost intervals. In other words, innermost intervals are the intervals that briefly summarize the total interval-valued data, and consist of combinations of elements

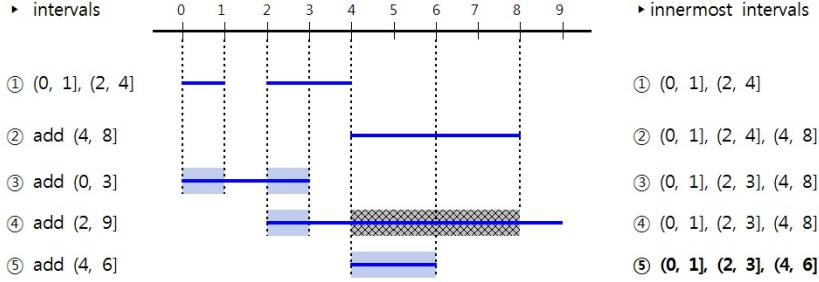


Figure 2.1: Innermost intervals.

of the sets  $\mathcal{L}$  and  $\mathcal{U}$ , respectively. When several intervals overlap each other, the smallest interval included in all of them becomes the innermost interval. On the other hand, when all intervals are disjoint each other, the set of intervals itself becomes the set of basis intervals and innermost intervals.

In our paper, we restrict our interest to the SCE based on innermost intervals among many. Finally, the recursion induced from (2.8) for the SCE,

$$w_j^{(r+1)} = \frac{1}{n} \sum_{i=1}^n \frac{\alpha'_{ij} w_j^{(r)}}{\sum_{l=1}^M \alpha'_{il} w_l^{(r)}}, \quad 1 \leq j \leq M \quad (2.9)$$

is defined with innermost intervals where  $\alpha'_{ij} = \mathbf{I}((t_j, s_j] \subseteq I_i)$ .

The recursive equation (2.9) finds the weights on the innermost intervals based on the observed intervals  $\{I_i = (\ell_i, u_i], i = 1, \dots, n\}$  and has the same form with the expectation-maximization algorithm for incomplete multinomial data (Tsai and Crowley, 1985). Multinomial distribution belongs to the exponential family. Thus Theorem 2 of Wu (1983) tells the convergence of the recursion in (2.9).

### 2.3.3. The almost sure limit of the SCE: self-consistent marginalization

We consider the almost sure (a.s.) limit of the SCE  $\widehat{f}_s$  or  $\widehat{F}_s$  (the cumulative distribution function) and, if it exists, define the self-consistent marginalization of interval-valued data as its a.s. limit.

The law of large numbers tells that the a.s. limit of the estimating equation (2.6) is

$$f(x) = E_I \left[ \frac{f(x)I(x \in I)}{\int_I f(t)dt} \right] = \int \left\{ \frac{f(x)I(x \in I)}{\int_I f(t) dt} \right\} dG(I), \quad (2.10)$$

where  $dG(I)$  is the probability distribution function of the interval  $I$ . We refer the solution to (2.10) as  $f_s$ , where  $\widehat{f}_s$  converges a.s. (Tsai and Crowley, 1985; Yu et al., 2000). Note that  $f_s$  (or  $F_s$ ) is also identifiable only up to the basis intervals of the population intervals, and we restrict the support of  $f_s$  (or  $F_s$ ) on innermost intervals like the SCE.

Our main question here is whether  $G(I)$  uniquely specify  $f_s$ . In other words, if we give  $G(I)$ , can we specify  $f_s$  and

$$\mathcal{F}(G) = \left\{ f \geq 0 \mid f(x) = \int \left\{ \frac{f(x)I(x \in I)}{\int_I f(t) dt} \right\} dG(I), \quad \int f(t)dt = 1 \right\}?$$

The answer is yes if  $G(I)$  has the support on a set of finite disjoint intervals, but it is not easy to specify  $f_s$  if some of the intervals in the support are overlapped. Below, we find this with a few examples. For simplicity, we assume that  $G$  has the support on a set of finitely many intervals. We can find the solution by solving a set of equations obtained by applying the integral equation (2.10) to each innermost interval. Two examples are the cases where inter-

vals are disjoint each other, and the other two are the cases where intervals overlap.

- (i) Example 1: Let  $dG(I)$  has a mass on only two disjoint intervals. That is,

$$dG(I) = \begin{cases} p & \text{if } I = (-\infty, c], \\ 1 - p & \text{if } I = (c, \infty). \end{cases}$$

Note that two intervals  $(-\infty, c], (c, \infty)$  are the basis intervals and innermost intervals for  $G$ . Then for  $x \in (-\infty, c]$ , the integral equation is

$$f(x) = \int \left\{ \frac{f(t) \mathbf{I}(x \in I)}{\int_I f(t) dt} \right\} dG(I) = \frac{f(x)}{F(c)} p$$

and for  $x \in (c, \infty)$ , that is

$$f(x) = \int \left\{ \frac{f(t) \mathbf{I}(x \in I)}{\int_I f(t) dt} \right\} dG(I) = \frac{f(x)}{1 - F(c)} (1 - p).$$

Thus any  $f_s$  which satisfies  $F_s(c) = p$  is the solution.

- (ii) Example 2 : Let  $dG(I)$  has a mass on finitely many disjoint intervals. That is,

$$dG(I) = \begin{cases} p_0 & \text{if } I = (-\infty, c_1], \\ p_1 & \text{if } I = (c_1, c_2], \\ \dots & \\ p_m & \text{if } I = (c_m, \infty), \end{cases}$$

with  $\sum_{k=0}^m p_k = 1$ . Then similarly to case 1, any  $f_s$  which satisfies  $F_s(c_{k+1}) - F_s(c_k) = p_k$ ,  $k = 0, \dots, m$  (where  $c_0 = -\infty$

and  $c_{m+1} = \infty$ ) is the solution. Note that the probability of  $I$  measured by  $G$  is preserved on a real line where the measure is  $F_s$ . That is,  $dG(I_k = (c_k, c_{k+1}]) = p_k = F_s(c_{k+1}) - F_s(c_k)$ .

Now we consider the cases where intervals overlap.

- (iii) Example 3: Let  $dG(I)$  has a mass on only two overlapping intervals. That is,

$$dG(I) = \begin{cases} p & \text{if } I = (\ell_1, u_1], \\ 1 - p & \text{if } I = (\ell_2, u_2], \end{cases}$$

where  $\ell_1 < \ell_2 < u_1 < u_2$  and  $p > 0$ . Nontrivial basis intervals are  $(\ell_1, \ell_2]$ ,  $(\ell_2, u_1]$ , and  $(u_1, u_2]$  and only  $(\ell_2, u_1]$  makes up the innermost interval. Thus the solution  $f_s$  has a mass only on  $(\ell_2, u_1]$ : that is,  $f_s(x)$  which satisfies  $F_s(u_1) - F_s(\ell_2) = 1$  and  $f_s(x) = 0$  for  $x \notin (\ell_2, u_1]$  is the solution.

- (iv) Example 4: We consider the case  $dG(I)$  has a mass on following three intervals.

$$dG(I) = \begin{cases} p & \text{if } I = (\ell_1, u_1], \\ 2p & \text{if } I = (\ell_2, u_2], \\ 1 - 3p & \text{if } I = (\ell_3, u_3], \end{cases}$$

where  $\ell_1 = \ell_2 < u_1 = \ell_3 < u_2 < u_3$  and  $p > 0$ . Then nontrivial basis intervals are three:  $(\ell_1, u_1]$ ,  $(\ell_3, u_2]$ ,  $(u_2, u_3]$ . Note that two intervals,  $(\ell_1, u_1]$  and  $(\ell_3, u_2]$  constitute the innermost

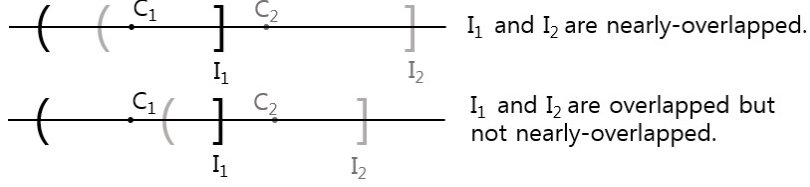


Figure 2.2: Nearly-overlapped intervals.

intervals. The integral equation is

$$f(x) = \begin{cases} \frac{f(x)}{\int_{\ell_1}^{u_1} f(t) dt} p + \frac{f(x)}{\int_{\ell_2}^{u_2} f(t) dt} 2p & \text{if } x \in (\ell_1, u_1], \\ \frac{f(x)}{\int_{\ell_2}^{u_2} f(t) dt} 2p + \frac{f(x)}{\int_{\ell_3}^{u_3} f(t) dt} (1 - 3p) & \text{if } x \in (\ell_3, u_2], \\ 0 & \text{otherwise.} \end{cases}$$

By solving the equation, we get the solution  $f_s(x)$  which satisfies  $F_s(u_1) - F_s(\ell_1) = \frac{p}{1-2p}$ ,  $F_s(u_2) - F_s(\ell_3) = \frac{1-3p}{1-2p}$ , and  $f_s(x) = 0$  for  $x \notin (\ell_1, u_1] \cup (\ell_3, u_2]$ .

On the other hand, if interval-valued data satisfies some conditions,  $f_s$  can be specified as  $f_C$ , the probability density function of the center of the interval. Before introducing the conditions, we define some further notations.

**Definition 2.1.** For the interval  $I_i$  and  $I_j$ , we say that the two intervals are nearly-overlapped if  $I_i \cap I_j$  includes the center of the interval  $I_i$  or  $I_j$ .

Figure 2.2 shows the graphical illustration of the nearly-overlapped intervals.

To define the concepts related to the center of the interval more conveniently, we use the C-R representation for interval-valued

data. That is, any interval-valued data  $I = (L, U]$  can be represented by  $(C - R, C + R]$ , where  $C$  and  $R$  denote the center and half-range of the interval, respectively. That is,  $C = (L + U)/2$  and  $R = (U - L)/2$ . Suppose that the center and range take finitely many values. Let  $C$  take values of the set  $\{c_1, \dots, c_A\}$  ( $A < \infty$ ) and  $R$  take values of the set  $\{r_1, \dots, r_B\}$  ( $B < \infty$ ). We sort the values of the range so that  $r_1 < \dots < r_B$ . Then, any interval  $I_j$ ,  $j = 1, \dots, N_G$  ( $N_G \leq AB$ ) can be represented by  $(c_a - r_b, c_a + r_b]$  for some  $a$  and  $b$ , where  $1 \leq a \leq A$ ,  $1 \leq b \leq B$ . We denote the interval  $(c_a - r_b, c_a + r_b]$  as  $\mathcal{I}_{ab}$ . The set of intervals which have center  $c_a$  in common is denoted by  $\mathcal{I}_{a+}$ . That is,  $\mathcal{I}_{a+} = \{\mathcal{I}_{ak} \mid k \in \{1, \dots, B\}\}$ .

Now, we say that the center  $c_a$  is *isolated* if  $\mathcal{I}_{a+} \cap \mathcal{I}_{k+} = \emptyset$  for all  $k \neq a$ . We say that the centers  $c_a$  and  $c_k$  are *bi-paired* each other if (i)  $\mathcal{I}_{a+} \cap \mathcal{I}_{j+} = \emptyset$  for all  $j \in \{1, \dots, B\} \setminus \{a, k\}$ , (ii)  $\mathcal{I}_{k+} \cap \mathcal{I}_{j+} = \emptyset$  for all  $j \in \{1, \dots, B\} \setminus \{k, a\}$ , and (iii) there exists  $m$ ,  $1 \leq m \leq B$  such that  $\mathcal{I}_{aj} \cap \mathcal{I}_{kj} = \emptyset$  for all  $j < m$  and  $\mathcal{I}_{aj}$  and  $\mathcal{I}_{kj}$  are nearly-overlapped each other for  $j \geq m$ .

Finally, we describe the lemma on the case  $f_s(x)$  equals to  $f_c(x)$ , in the sense that  $f_s(x)$  and  $f_c(x)$  assign the same mass on every innermost interval.

**Lemma 2.1.** *Assume that the population of interval-valued data has a support on a set of finitely many intervals and whose centers are isolated or bi-paired. If the center and range are independent of each other, the probability density function of the center denoted by  $f_c(x)$  is the solution of the integral equation (2.10).*

*Proof.* Since we assume that the support is finite, the center and range take finitely many values. Let the values that  $C$  can take are



$c_1, \dots, c_A$ , for some interger  $A$  and the values that  $R$  can take are  $r_1, \dots, r_B$ , for some interger  $B$ . We sort the values of the ranges so that  $r_1 < \dots < r_B$ . Then, the interval  $I$  can take  $AB$  number of intervals under the assumption of independence of the center and range. Thus, for any interval  $\mathbb{I}_{ab} = (c_a - r_b, c_a + r_b]$ ,  $a = 1, \dots, A$ ,  $b = 1, \dots, B$ ,

$$dG_I(\mathbb{I}_{ab}) = Pr(I = \mathbb{I}_{ab}) = Pr(C = c_a) \cdot Pr(R = r_b)$$

holds. To simplify the notation,  $Pr(C = c_a)$  is denoted by  $P_C(c_a)$  and  $Pr(R = r_b)$  is denoted by  $P_R(r_b)$ .

Under the assumptions above, innermost intervals includes all the center points and let us denote the inner most interval includes the center  $c_a$  as  $inn(c_a)$  for  $a = 1, \dots, A$ . Then the innermost intervals  $inn(c_a)$  has two types:  $\mathbb{I}_{a1} = (c_a - r_1, c_a + r_1]$  or an interval containing  $c_a$  as a subset of  $\mathbb{I}_{a1}$ .

Now, first suppose  $x \in inn(c_a)$  for some  $a$ ,

- (i) If the center  $c_a$  is isolated.

The innermost interval containing  $c_a$  is naturally  $\mathbb{I}_{a1}$ . Thus, for  $x \in \mathbb{I}_{a1}$ , substituting the probability density function of the center  $f_C(x)$  into the integral equation (2.10),

$$\begin{aligned} \int \left\{ \frac{f_C(x) \mathbb{I}(x \in I)}{\int_I f_C(t) dt} \right\} dG(I) &= \sum_{b=1}^B \frac{f_C(x)}{\int_{c_a - r_b}^{c_a + r_b} f_C(t) dt} Pr(I = \mathbb{I}_{ab}) \\ &= \sum_{b=1}^B \frac{f_C(x)}{\int_{c_a - r_b}^{c_a + r_b} f_C(t) dt} P_C(c_a) P_R(r_b) \text{ by indep.} \\ &= \sum_{b=1}^B \frac{f_C(x)}{P_C(c_a)} P_C(c_a) P_R(r_b) = f_C(x). \end{aligned}$$

- (ii) If the center point  $c_a$  is bi-paired with the center point  $c_k$ ,  $k \neq a$ . Let  $m$ ,  $1 \leq m \leq B$  be the smallest number such that

$c_a \in \mathbf{I}_{kb}$ , for all  $b \geq m$ . Then for  $b \geq m$ ,  $c_k \in \mathbf{I}_{ab}$  also holds by symmetry. The integral equation (2.10) with  $f_C(x)$  is,

$$\begin{aligned}
& \int \left\{ \frac{f_C(x) \mathbf{I}(x \in I)}{\int_I f_C(t) dt} \right\} dG(I) \\
&= \sum_{b=1}^B \frac{f_C(x)}{\int_{c_a-r_b}^{c_a+r_b} f_C(t) dt} Pr(I = \mathbf{I}_{ab}) + \sum_{b=m}^B \frac{f_C(x)}{\int_{c_k-r_b}^{c_k+r_b} f_C(t) dt} Pr(I = \mathbf{I}_{kb}) \\
&= \sum_{b=1}^{m-1} \frac{f_C(x)}{\int_{c_a-r_b}^{c_a+r_b} f_C(t) dt} P_C(c_a) P_R(r_b) + \sum_{b=m}^B \frac{f_C(x)}{\int_{c_a-r_b}^{c_a+r_b} f_C(t) dt} P_C(c_a) P_R(r_b) \\
&\quad + \sum_{b=m}^B \frac{f_C(x)}{\int_{c_k-r_b}^{c_k+r_b} f_C(t) dt} P_C(c_k) P_R(r_b) \\
&= \sum_{b=1}^{m-1} f_C(x) P_R(r_b) + \sum_{b=m}^B \frac{f_C(x)}{P_C(c_a) + P_C(c_k)} P_C(c_a) P_R(r_b) \\
&\quad + \sum_{b=m}^B \frac{f_C(x)}{P_C(c_k) + P_C(c_a)} P_C(c_k) P_R(r_b) \\
&= \sum_{b=1}^{m-1} f_C(x) P_R(r_b) + \sum_{b=m}^B f_C(x) P_R(r_b) = f_C(x).
\end{aligned}$$

Second, suppose  $x \notin \text{inn}(c_a)$  for all  $a = 1, \dots, A$ , then  $f_C(x) = 0$  and the integral equation holds.

□

## 2.4. Numerical study

In this section, we compare the SCE to the two existing estimators, uniform kernel estimator (UK) and Gaussian kernel estimator (GK) in various settings. Also, we investigate the sensitivity of innermost intervals of the SCE depending on how data is rounded to decimal places.

We use the following notation:  $F_C$  represents the cumulative

distribution function (c.d.f.) of the center of the interval,  $F_A$  represents the c.d.f. of the marginal aggregation (or histogrammed marginalization), and  $F_s$  represents the c.d.f. of the self-consistent marginalization, the a.s. limit of the SCE in Section 2.3. To approximate  $F_A$ , we generate 10,000 intervals according to the simulation setting and evaluate the estimator as its limit following Jeon et al. (2015). When the explicit form of  $F_s$  is not available, we also approximate the limit using 10,000 intervals. We set the sample size  $n$  of a single data set of each case as 100 and replicate 100 data sets to compare the performance of estimators.

#### 2.4.1. Continuous cases

Let the center  $C_i = \frac{L_i + U_i}{2}$  and the half-range  $R_i = \frac{U_i - L_i}{2}$ , for  $i = 1, 2, \dots, n$  with  $n = 100$ . We consider two cases below, where one assumes  $C_i$  and  $R_i$  are independent of each other and the other assumes they are dependent. We refer to the first case as (C.IND) and the second as (C.DEP). About the naming the cases, C represent  $dG(I)$  is continuous, and the three characters after the dot(.) indicate the relationship of  $C$  and  $R$ , where IND indicates “independent” and DEP for “dependent”. These two cases refer to the cases in Jeon et al. (2015), and in each case we extended the variation of the range.

- (C.IND) The center  $C_i$  is generated from  $N(5, 2^2)$  and the half-range  $R_i$  is independently generated from  $U(2.5, 3.5)$ .
- (C.DEP) The center  $C_i$  is generated from  $N(5, 3^2)$  and the half-range  $R_i$  has linear relationship with the center:  $R_i = 4 - 0.2C_i + \tau_i$ ,  $\tau_i \sim N(0, 1)$ .

We use three distance measures to compare the estimators:  $\ell_1$ ,  $\ell_2$  and  $\ell_\infty$ . The  $\ell_1$  measure is defined as the mean of absolute difference between the estimated distribution ( $\widehat{F}$ ) and target distribution ( $F_T$ ),  $\frac{1}{M} \sum_{m=1}^M |\widehat{F}(x_m) - F_T(x_m)|$ . The  $\ell_2$  measure is defined as  $\sqrt{\frac{1}{M} \sum_{m=1}^M \{\widehat{F}(x_m) - F_T(x_m)\}^2}$ . The  $\ell_\infty$  measure is defined as  $\max_m |\widehat{F}(x_m) - F_T(x_m)|$ , where  $x_1, \dots, x_M$  are the evenly spaced grid points on the pre-decided domain. Each domain is designed to cover approximately 95% of the support of the interval considering the distribution of the interval. The number of grid points,  $M$  is set to be 100 for each case. For the grid points where  $\widehat{F}$  is not specified, we linearly interpolate them.

First, we find from Figure 2.3 that  $F_C$ ,  $F_s$ , and  $F_A$  are different from one another. If we define the distance between two distribution functions  $F$  and  $G$  as  $\sup_x |F(x) - G(x)|$ , the distance between  $F_s$  and  $F_A$  is larger than the other two distances between the three

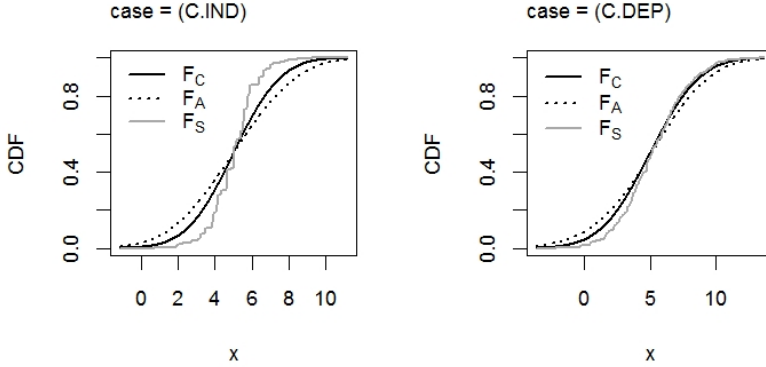


Figure 2.3: Display of the marginal distribution functions  $F_C$ ,  $F_A$  and  $F_s$  for each case.

Table 2.1: Simulation results for continous cases. The distance between the estimated distribution ( $\widehat{F}(x)$ ) and target distribution ( $F_T(x)$  with  $T = "C", "A", "S"$ ) is displayed with the standard errors in parentheses.

estimation method			SCE		UK		GK	
(C.IND)	$F_C$	$\ell_1$	0.085	(0.015)	0.052	(0.007)	0.048	(0.008)
		$\ell_2$	0.111	(0.021)	0.058	(0.009)	0.053	(0.010)
		$\ell_\infty$	0.281	(0.057)	0.093	(0.017)	0.085	(0.018)
	$F_A$	$\ell_1$	0.131	(0.017)	0.017	(0.010)	0.018	(0.010)
		$\ell_2$	0.152	(0.020)	0.019	(0.011)	0.020	(0.011)
		$\ell_\infty$	0.317	(0.048)	0.030	(0.016)	0.032	(0.016)
	$F_s$	$\ell_1$	0.044	(0.009)	0.118	(0.007)	0.114	(0.008)
		$\ell_2$	0.076	(0.018)	0.134	(0.007)	0.129	(0.008)
		$\ell_\infty$	0.270	(0.067)	0.243	(0.018)	0.237	(0.019)
(C.DEP)	$F_C$	$\ell_1$	0.051	(0.009)	0.037	(0.010)	0.036	(0.010)
		$\ell_2$	0.066	(0.013)	0.042	(0.012)	0.041	(0.012)
		$\ell_\infty$	0.185	(0.038)	0.071	(0.020)	0.066	(0.021)
	$F_A$	$\ell_1$	0.073	(0.013)	0.020	(0.010)	0.020	(0.010)
		$\ell_2$	0.087	(0.014)	0.024	(0.012)	0.023	(0.012)
		$\ell_\infty$	0.204	(0.035)	0.044	(0.020)	0.041	(0.019)
	$F_s$	$\ell_1$	0.041	(0.009)	0.062	(0.010)	0.061	(0.010)
		$\ell_2$	0.058	(0.013)	0.071	(0.012)	0.070	(0.012)
		$\ell_\infty$	0.179	(0.046)	0.129	(0.023)	0.127	(0.023)

distributions. Second, Table 2.1 shows that the UK and GK are almost identical as expected and that the two estimators estimate  $F_A$  very well. The SCE estimates  $F_s$  better than the UK and GK based on the  $\ell_1$  and  $\ell_2$  metrics.

### 2.4.2. Discrete cases

We now consider the cases where  $dG(I)$  is discrete. We consider three cases, where the first case has a support on a set of finite disjoint intervals and the other two cases have supports on the overlapping intervals. We denote the first case as (D.DIS) where D means “discrete” and DIS represents “disjoint”, and the other two cases are denoted by (D.OVL1) and (D.OVL2) where OVL represents “overlapped”.

For the cases with the support on finitely many intervals, we assume that there are  $K$  basis intervals. Then the targets are  $\mathbf{p}_s = (p_{s_1}, \dots, p_{s_K})$  and  $\mathbf{p}_A = (p_{A_1}, \dots, p_{A_K})$ , where  $p_{s_k}$  and  $p_{A_k}$ ,  $k = 1, \dots, K$  are the probabilities for the  $k$ -th basis interval measured by  $F_s$  and  $F_A$ , respectively. Recall that innermost intervals are the subset of basis intervals, and  $\mathbf{p}_s$  has a positive mass only on a set of innermost intervals. In each case, we compare the  $\mathbf{p}_s$  (or  $\mathbf{p}_A$ ) with the estimated  $\mathbf{p}$ , notated as  $\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_K)$  where  $\hat{p}_k = \hat{F}(u_k) - \hat{F}(\ell_k)$ , with the three distance metrics  $\ell_1$ ,  $\ell_2$ , and  $\ell_\infty$ :  $\ell_1 = \frac{1}{K} \sum_{k=1}^K |\hat{p}_k - p_{T_k}|$ ,  $\ell_2 = \sqrt{\frac{1}{K} \sum_{k=1}^K (\hat{p}_k - p_{T_k})^2}$ , and  $\ell_\infty = \max_k |\hat{p}_k - p_{T_k}|$ , where  $p_{T_k}$  is  $p_{s_k}$  or  $p_{A_k}$ .

Note that, when  $dG(I)$  is discrete, the SCE and UK provide a discrete summary for the basis intervals, whereas the GK provides a continuous density function on a real line. The probabilities for the basis intervals by the GK are evaluated from this continuous density.

(D.DIS) Assume  $dG(I)$  has a mass on  $\{(0, 1], (1, 2], \dots, (7, 8]\}$ . Note that the intervals are disjoint each other. Recall that when the support of the intervals is a set of finite disjoint intervals,

the support itself becomes the set of basis intervals and innermost intervals. Let the support be  $\{I_1, \dots, I_K\}$ , then as we mentioned in Section 2.3,  $p_{s_k} = dG(I_k)$ ,  $k = 1, \dots, K$  holds. Also, naturally  $\mathbf{p}_A = \mathbf{p}_S$  holds, and the SCE and UK are exactly the same. As an example, the probabilities,  $\mathbf{p}_A$  and  $\mathbf{p}_S$  for the interval  $I_k = (\ell_k, u_k]$ ,  $k = 1, \dots, K$  are represented in Table 2.2.

Table 2.2:  $dG(I)$ ,  $\mathbf{p}_S$  and  $\mathbf{p}_A$  for (D.DIS).

case	$dG(I) = \mathbf{p}_S = \mathbf{p}_A$							
	(0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	(5, 6]	(6, 7]	(7, 8]
(D.DIS)	1/40	4/40	6/40	5/40	9/40	8/40	5/40	2/40

(D.OVL) We use 4 intervals  $(0, 1], (1, 3], (0, 2], (2, 3]$  as the support following an example of Gentleman and Geyer (1994). Both the basis intervals and innermost intervals for this case are  $(0, 1]$ ,  $(1, 2]$ , and  $(2, 3]$ . We consider 2 different choices for  $dG(I)$  listed in Table 2.3. For each case, we can find  $\mathbf{p}_S$  by solving the integral equation (2.10) and  $\mathbf{p}_A$  is also easily calculated.

Table 2.3:  $dG(I)$  and the corresponding  $\mathbf{p}_S$  and  $\mathbf{p}_A$  for cases (D.OVL1) and (D.OVL2).

case	$dG(I)$				$\mathbf{p}_S$			$\mathbf{p}_A$		
	$I = (0, 1]$	(1, 3]	(0, 2]	(2, 3]	(0, 1]	(1, 2]	(2, 3]	(0, 1]	(1, 2]	(2, 3]
(D.OVL1)	2/9	2/9	4/9	1/9	0.5	0.3	0.2	4/9	3/9	2/9
(D.OVL2)	1/3	1/6	1/3	1/6	2/3	0	1/3	2/4	1/4	1/4

Table 2.4 reports the three distance measures between the esti-

Table 2.4: Simulation results for discrete case. The estimation errors with their standard errors in parentheses.

			SCE		UK		GK	
(D.DIS)	$\mathbf{p}_s (= \mathbf{p}_A)$	$\ell_1$	0.025	(0.006)	0.025	(0.006)	0.023	(0.006)
		$\ell_2$	0.031	(0.008)	0.031	(0.008)	0.028	(0.007)
		$\ell_\infty$	0.059	(0.019)	0.059	(0.019)	0.054	(0.017)
(D.OVL1)	$\mathbf{p}_A$	$\ell_1$	0.068	(0.038)	0.027	(0.016)	0.027	(0.015)
		$\ell_2$	0.077	(0.043)	0.030	(0.018)	0.032	(0.016)
		$\ell_\infty$	0.102	(0.056)	0.041	(0.025)	0.046	(0.022)
	$\mathbf{p}_s$	$\ell_1$	0.060	(0.034)	0.042	(0.022)	0.046	(0.020)
		$\ell_2$	0.067	(0.038)	0.047	(0.024)	0.055	(0.023)
		$\ell_\infty$	0.090	(0.0503)	0.064	(0.033)	0.084	(0.033)
(D.OVL2)	$\mathbf{p}_A$	$\ell_1$	0.145	(0.029)	0.030	(0.018)	0.030	(0.015)
		$\ell_2$	0.161	(0.033)	0.034	(0.02)	0.036	(0.017)
		$\ell_\infty$	0.218	(0.043)	0.046	(0.027)	0.055	(0.026)
	$\mathbf{p}_s$	$\ell_1$	0.047	(0.031)	0.167	(0.016)	0.180	(0.013)
		$\ell_2$	0.053	(0.035)	0.183	(0.019)	0.194	(0.016)
		$\ell_\infty$	0.070	(0.047)	0.250	(0.024)	0.251	(0.021)

mates and targets. In case (D.DIS) where the intervals are disjoint each other, both the SCE (=UK) and GK show similar performance in estimating  $\mathbf{p}_s (= \mathbf{p}_A)$ . Figure 2.4 displays the box plots of the difference between the estimated and target probabilities for the interval  $I_k, k = 1, 2, \dots, K$  and density functions. The SCE and GK perform similarly and well together. The density plot confirms that the GK provides the continuous density function on a real line while the SCE and UK provide only the discrete summary for the basis intervals. In cases (D.OVL1) and (D.OVL2), it is also clear that the GK and UK estimate  $\mathbf{p}_A$  and SCE estimates  $\mathbf{p}_s$ . Note that in case (D.OVL2), the SCE (or self-consistent marginaliza-



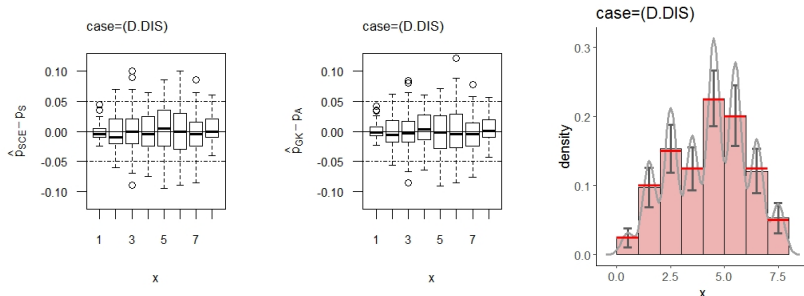


Figure 2.4: Boxplots and density plots for (D.DIS). The third plot represents the estimated probability density functions. The height of the bar represents the average probability that the SCE(=UK) estimates for each interval for the 100 data sets. The standard error bar is also plotted. The red line indicates the true probability  $p_s (= p_A)$ . The gray line is the average of the probability density function by the GK for the 100 data sets.

tion) has no mass at the interval  $(1, 2]$  whereas the UK and GK (or histogrammed marginalization) have a positive mass at  $(1, 2]$ . Thus, in case (D.OVL2), the numeric error of SCE to  $p_s$  is much smaller than those of other estimates to  $p_s$ . In addition, when focusing on each basis interval, Figure 2.5 shows that the GK has a bias in estimating  $p_A$  at the interval  $(0, 1]$  compared to the UK.

### 2.4.3. The sensitivity to the coarseness of the intervals

Unlike the UK and GK, the SCE finds a succinct representation to data and is based on innermost intervals. One question we may have is how the innermost intervals and SCE are sensitive to small changes in observed intervals. To understand this, we examine

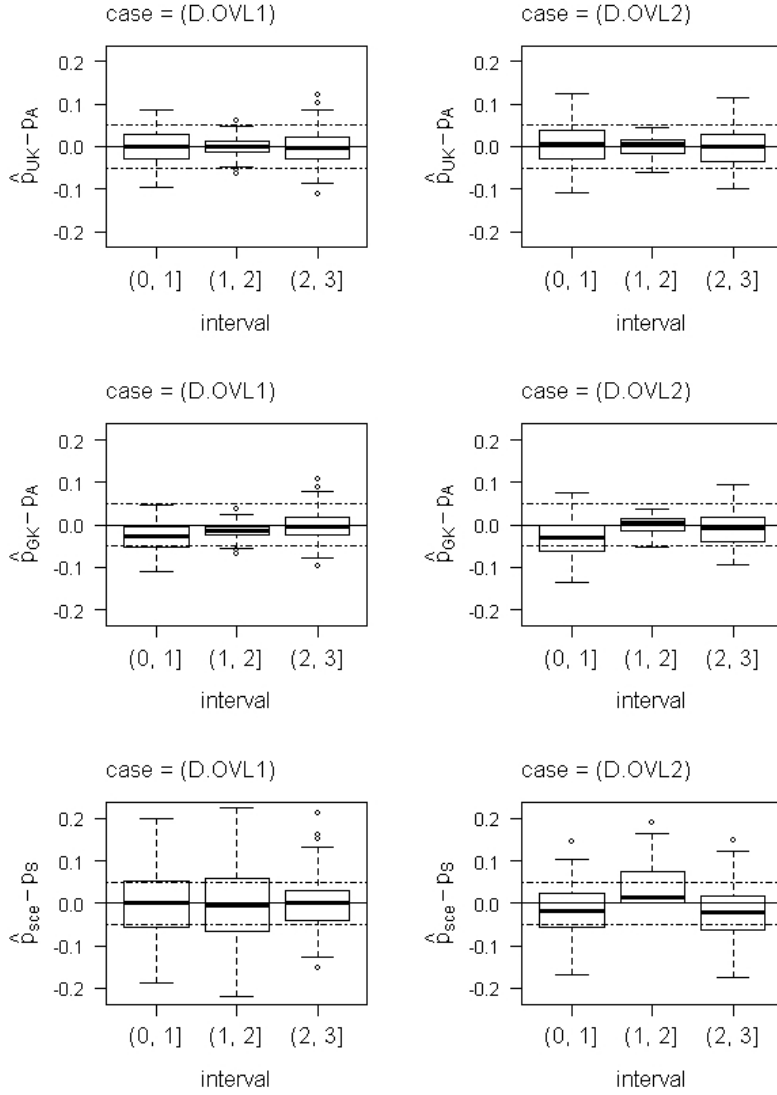


Figure 2.5: Boxplots for (D.OVL1) and (D.OVL2).

how the innermost intervals and SCE vary with the degrees of coarseness (or precisions) of the intervals.

The study design is as follows. We first generate raw intervals (Ori) from a continuous distribution and round them to the first (R1) and second (R2) decimal places. We then compare the innermost intervals and SCE in each case. The original data are generated from the above case (C.IND).

Table 2.5 shows raw data generated from the setting and the data rounded to the first and second decimal places. Notice that as data is rounded to lower digits, the total number of unique upper and lower bounds of interval-valued data also decreases. In particular, when continuous raw data is rounded to the first digit, the number of unique upper and lower bounds is quickly reduced from 200 to 99, more than half.

Table 2.6 shows the innermost intervals and corresponding SCEs in each case. Note that as data are rounded to lower digits, the number of innermost intervals also decreases. First, look at

Table 2.5: Data for each case.

obs. index	(Ori)	(R2)	(R1)
1	(3.1916, 9.1194]	(3.19, 9.12]	(3.2, 9.1]
2	(0.9791, 7.0086]	(0.98, 7.01]	(1.0, 7.0]
3	(-1.2625, 3.9636]	(-1.26, 3.96]	(-1.3, 4.0]
4	(-1.1008, 4.4074]	(-1.10, 4.41]	(-1.1, 4.4]
$\vdots$	$\vdots$	$\vdots$	$\vdots$
100	(2.5592, 9.2654]	(2.56, 9.27]	(2.6, 9.3]
# of unique $\{\ell_i, u_i\}_{i=1}^{100}$	200	180	99

Table 2.6: The innermost intervals and SCE for each case. The numbers in parentheses([ ]) to the right of case notation indicate the total number of innermost intervals.

index	(Ori) [9]		(R2) [8]		(R1) [7]	
	intervals	$\hat{p}$ (SCE)	intervals	$\hat{p}$ (SCE)	intervals	$\hat{p}$ (SCE)
1	(3.1806, 3.1910]	0.143	(3.18, 3.19]	0.143	(3.0, 3.2]	0.125
2	(3.5482, 3.5807]	0.023	(3.55, 3.58]	0.023	(3.5, 3.6]	0.041
3	(3.9217, 3.9636]	0.035	(3.92, 3.96]	0.035	(3.9, 4.0]	0.035
4	(4.1901, 4.2686]	0.355	(4.19, 4.27]	0.355	(4.2, 4.3]	0.355
5	(4.3877, 4.4074]	0.000	(4.39, 4.41]	0.000		
6	(4.9094, 4.9506]	0.244	(4.91, 4.95]	0.244	(4.9, 5.0]	0.244
7	(5.2366, 5.2426]	0.000				
8	(5.3147, 5.4671]	0.129	(5.31, 5.47]	0.129	(5.3, 5.5]	0.129
9	(6.8000, 6.8696]	0.071	(6.80, 6.87]	0.071	(6.8, 6.9]	0.071

cases (Ori) and (R2). Except that the 7–th interval of case (Ori) is missing in case (R2), all other intervals are the same each other if we represent the intervals in case (Ori) only to the second decimal point. The SCEs of cases (Ori) and (R2) are the exactly same each other as the 7–th interval in case (Ori) has no mass.

Meanwhile, case (R1) shows a different pattern. In particular, the first innermost interval (3.1806,3.1910] of case (Ori) is not preserved in case (R1). Rounding the interval (3.1806,3.1910] to the first decimal point yields (3.2,3.2], which is an empty set since the upper and lower values are the same. The innermost interval at case (R1) corresponding to the interval (3.1806,3.1910] is (3.0,3.2], which is broader than that of case (Ori) (or (R2)). This is because the order of original data is not preserved as data is rounded. If the order of original data is not preserved, the over-

lapping structure of the intervals changes: therefore, the innermost intervals also vary.

Specifically, the intervals that directly affect the construction of the first innermost interval in original data are  $(-1.8211, 3.1910]$  and  $(3.1806, 8.7454]$ . These two intervals overlap, and the common interval between the two is  $(3.1806, 3.1910]$ , which is the first innermost interval in case (Ori). However, rounding to the first decimal place yields  $(-1.8, 3.2]$  and  $(3.2, 8.7]$ , respectively, and these two are disjoint. That is, two overlapping intervals become disjoint by rounding, which causes the innermost interval to change. The other innermost intervals in case (R1) are the same as the corresponding innermost intervals of cases (Ori) and (R2) since the rounding preserves the order of upper and lower values of the original intervals.

Therefore, the mass estimate for the first innermost interval  $(3.0, 3.2]$  in case (R1) is different from that in case (Ori) (or (R2)). Notice that as the mass difference in the first innermost interval is reflected in the second innermost interval, the sum of the masses assigned to the first and second intervals is the same to that of case (Ori) (or (R2)). The SCEs in the other intervals are the same in all cases except for the first and second innermost intervals.

In short, if a small variation to existing data changes the overlapping structure of interval-valued data, the innermost intervals and SCE also vary accordingly.

## 2.5. Data examples

We apply the marginalization methods discussed in this chapter to two real data examples. One is the rally data that records the arrival and departure times of individual participants, and the other is the blood pressure data consisting of diastolic and systolic blood pressure.

### 2.5.1. Rally data

The rally data is the survey from Korea’s March for Science, a global event across the world on Earth day (April 22) in 2017. In the survey, the rally participants answered when they arrived at and left the rally. That is, the response of one participant is an interval composed of arrival and departure times, which is the MM-type interval-valued data. Of the approximately 1,000 participants, 129 participants responded to questions about their participation time, and we use  $N = 125$  data, excluding 4 wrong answers. We apply the SCE and GK to the rally data to estimate the marginal distribution and look at the properties and implications of the distribution.

Figure 2.6 shows the interval plot of data and the SCE and GK based on this data. The SCE provides a discrete summary concentrated on specific intervals, while the GK provides continuous density over a wide area. Table 2.7 is the SCE summary, which shows the innermost intervals and corresponding masses.

According to the Science News “Live updates from the global

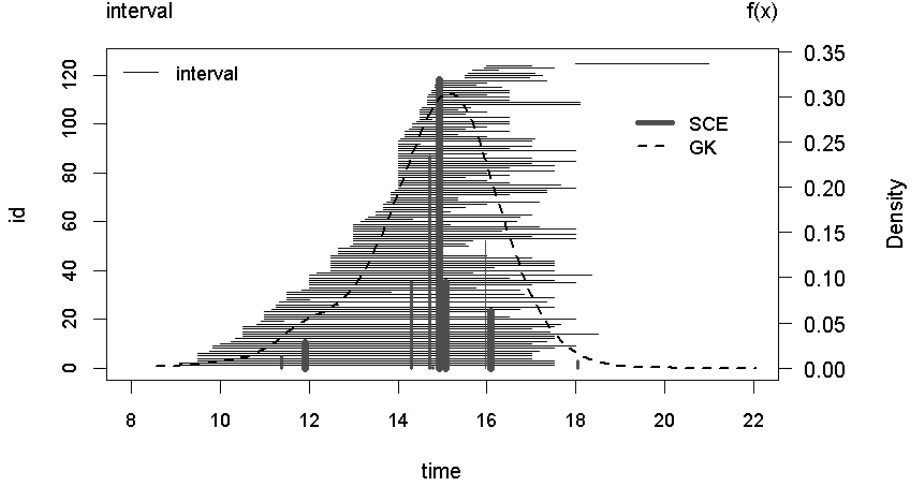


Figure 2.6: Rally data and the SCE and GK. In the SCE plot, the height of the bar indicates the size of the mass, and the width of the bar indicates the width of the innermost interval.

Table 2.7: Innermost intervals and corresponding SCE. Intervals with masses greater than 0.01 are displayed. The numbers in parentheses are the GK estimates.

interval	$\hat{p}$	interval	$\hat{p}$
(11:20, 11:25]	0.01 (0.00)	(15:00, 15:10]	0.09 (0.05)
(11:50, 12:00]	0.03 (0.01)	(15:57, 15:59]	0.14 (0.01)
(14:15, 14:20]	0.10 (0.02)	(16:00, 16:10]	0.06 (0.04)
(14:41, 14:45]	0.24 (0.02)	(18:00, 18:05]	0.01 (0.00)
(14:50, 15:00]	0.32 (0.05)	Total	1.00 (0.19)

March for Science”<sup>1</sup>, ten speeches were given starting at 2 p.m., and at 3 p.m., protestors marched from the Sejong Center to the Gwanghwamun district and returned to the center. The booths set up by science-related groups were open from 11 a.m. until 5 p.m..

Looking at the SCE, the innermost intervals where masses are concentrated seem to explain well the major events of the rally. Specifically, a mass of 0.10 is assigned on (14:15, 14:20], which is the time when the speech was going well. Notably, three intervals ((14:41, 14:45], (14:50, 15:00] and (15:00, 15:10]) around 3 p.m., when the march began, have a mass of 0.65 which is much more than half. The GK also shows a peak at 3:10, but there is only a mass of 0.14 on the continuous interval((14:41, 15:10]) where the SCE has a mass of 0.65. The next concentration with a mass of 0.20 appears at two close intervals at about 4 p.m., which is one hour after the march began and one hour before the booth closed. As a small concentration before the main events, the SCE has a positive mass on (11:50, 12:00], around one hour after opening the booths.

This shows that the SCE provides more concise information than the GK as the SCE assigns a weight on the representative intervals.

### **2.5.2. Blood pressure data**

We use the data from National Heart, Lung, and Blood Institute Growth and Health Study (NGHS), which is a cohort study to evaluate the temporal trends of cardiovascular risk factors, such

---

<sup>1</sup><https://www.sciencemag.org/news/2017/04/live-updates-global-march-science>



as systolic and diastolic blood pressures (SBP, DBP) based on up to ten annual visits of 2,379 African-American and Caucasian girls. The blood pressure (BP) data, which is measured at two levels, is also an example of the MM-type interval-valued data, as mentioned in the introduction. We use the BP data from the first visit and estimate its marginal distribution. We remove subjects with a missing measurement in either SBP or DBP. The total number of subjects is  $N = 2,256$  and mean of the center of BP (mid-BP) is 79.41. The goal of this study is to find the difference in BP records between African-American and Caucasian girls if any.

Table 2.8 shows descriptive statistics of the BP data by race and the test (t-test) results on whether the BP of African-Americans is greater than that of Caucasians. In all three BPs, mid-BP, SBP and DBP, mean value of African-American girls found to be higher than that of Caucasians.

To understand its distribution beyond the mean better, we apply the SCE, UK and GK to estimate the marginal distribution

Table 2.8: Descriptive statistics of the BP data by race. The p-value is from the t-test on the alternative hypothesis that the BP of African-American is higher than that of Caucasian. At the first column, mid-BP indicates the center of the BP data.

	Caucasian	African-American	p-value
mid-BP	78.67 (9.09)	80.13 (8.03)	< 0.0001
DBP	56.72 (12.19)	58.03 (11.72)	0.0047
SBP	100.62 (9.28)	102.23 (8.65)	< 0.0001

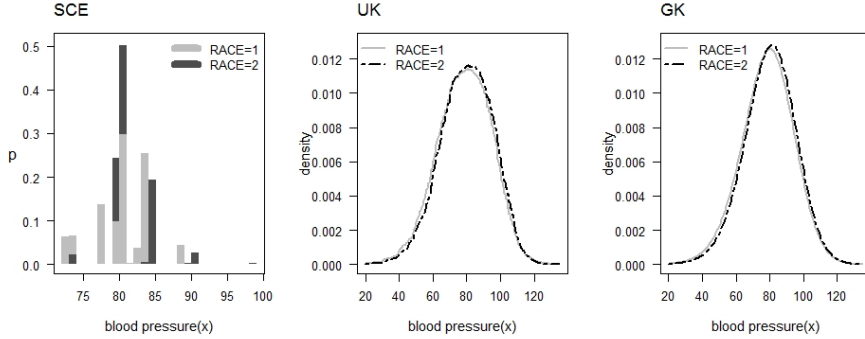


Figure 2.7: Density plots by RACE : Race = 1 indicates Caucasians, and RACE = 2 indicates African-Americans.

of the BP by race. The results are plotted in Figure 2.7. The SCE provides a discrete summary concentrated on specific intervals between 70 and 100, while the UK and GK provide (almost) continuous density over a wide area from 40 to 120. We find that the SCE shows the difference in BP between two races more clearly than the UK and GK. In particular, if we look into the innermost intervals which form the support of the SCE, they range from  $(72, 73]$  to  $(88, 89]$  for Caucasian, while those of African-American range from  $(73, 74]$  to  $(98, 99]$ . Thus, we conclude that the distribution of BP of the African-American has a longer right-tail than Caucasians.

## 2.6. Conclusion

In this chapter, we propose a new type of marginal distribution estimator (marginal summary), named as the SCE, for interval-valued data. We characterize the a.s. limit of the SCE as a new

type of marginal distribution of interval-valued data, named as a self-consistent marginalization. The self-consistent marginalization equals to the histogrammed marginal distribution when interval-valued data has the support on a finite number of disjoint intervals.

The numerical study and data examples show that the SCE (or self-consistent marginalization) provides a more concise (de-blurred) marginal representation to interval-valued data than the existing kernel-based methods, the UK and GK (or histogrammed marginalization). That is, the SCE gives a discrete data summary based on innermost intervals of the data for a single data set, often providing a better representation of the characteristics of the distribution, as in data examples.

We have looked at univariate interval-valued data on a real line so far. Similarly, we can extend the SCE to  $p$ -dimensional interval-valued data with almost no alteration. We conclude this chapter with a sketch of the extension and remain the details for the future study.

### 2.6.1. Extension to $p$ -dimensional interval-valued data

The data now consists of  $p$  variables, and each variable is interval-valued. Then each observation is represented by a  $p$ -dimensional hyper-rectangle in  $\mathcal{R}^p$ . Suppose we observe  $n$  independent  $p$ -dimensional hyper-rectangles  $\{R_i = (\ell_{1i}, u_{1i}] \times \cdots \times (\ell_{pi}, u_{pi}], i = 1, \dots, n\}$ . We define the SCE as follow similarly to the univariate case. The SCE, which we notate as  $\widehat{f}_s^p(\mathbf{x})$ ,  $\mathbf{x} \in \mathcal{R}^p$ , is the solution to the estimating equation

$$f^p(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \frac{f^p(\mathbf{x}) \mathbf{I}(\mathbf{x} \in R_i)}{\int_{R_i} f(\mathbf{t}) \, d^p \mathbf{t}}, \quad (2.11)$$

where  $\mathbf{x} = (x_1, \dots, x_p)$ ,  $\mathbf{t} = (t_1, \dots, t_p)$  and  $d^p \mathbf{t}$  is the  $p$ -dimensional volume differential.

We describe the maximal intersections corresponding to the innermost intervals in the univariate case following Gentleman and Vandal (2002) and Wong and Yu (1999).

First, we briefly introduce some basic concepts of graph theory and apply them to data. Let observed hyper-rectangle data be  $\mathbf{R} = \{R_1, \dots, R_n\}$ . Each observation  $R_i, i = 1, \dots, n$  corresponds to a vertex which will be denoted by its index. That is, the observation  $R_i$  will correspond to vertex  $i$ . The set of vertices is denoted  $\mathbf{V}$ . Two vertices  $j$  and  $k$  are joined by an edge if the corresponding  $R_j$  and  $R_k$  are intersect. The edge is denoted  $(j, k)$  and the set of edges is denoted  $\mathbf{E}$ . If  $(j, k) \in \mathbf{E}$ , then we say that  $j$  is adjacent to  $k$ . A graph  $\mathbf{G}$  is the collection of vertices and edges and written as  $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ . A clique  $\mathbf{C}$  is a subset of  $\mathbf{V}$  such that every two distinct vertices are adjacent. A maximal clique is a clique that cannot be a proper subset of any other clique.

Second, we form the intersection graph of data and get the maximal cliques. Every observation belongs to at least one maximal clique. Let the maximal cliques be denoted  $\mathcal{M} = \{M_1, \dots, M_m\}$ . Like the innermost intervals of the univariate case, the SCE in  $\mathcal{R}^p$  has the support on a finite number of disjoint hyper-rectangles, which is called the maximal intersections following Wong and Yu (1999). The maximal intersections  $\mathcal{H} = \{H_1, \dots, H_m\}$  are the real representations of the maximal cliques  $\mathcal{M}$  and defined as

$$H_j = \bigcap_{k \in M_j} R_k, \quad j = 1, \dots, m.$$

Finally, suppose  $\mathcal{H} = \{H_1, \dots, H_m\}$  is the maximal intersec-

tions from observed  $p$ -dimensional interval-valued data  $R_1, \dots, R_n$ . As in Theorem 1, we can show that the solution to (2.11) is only identifiable up to the maximal intersections  $H_1, \dots, H_m$  and (2.11) is simplified to

$$w_j = \frac{1}{n} \sum_{i=1}^n \frac{\alpha_{ji} w_j}{\sum_{l=1}^m \alpha_{li} w_l}, \quad 1 \leq j \leq m, \quad (2.12)$$

where

$$\alpha_{ji} = I(H_j \subseteq R_i) \quad \text{and} \quad w_j = \int_{H_j} f^p(\mathbf{x}) \, d^p \mathbf{x}.$$

We can further define the self-consistent marginalization of the  $p$ -dimensional interval distribution as the a.s. limit of the SCE defined above.

## Chapter 3

# Two-Sample Tests for Interval-Valued Data

### 3.1. Introduction

Among many statistical procedures, the comparison of two populations is one of the most fundamental statistical questions. However, little research is done for interval-valued data. The only method we aware is the combined test (CB) by Grzegorzewski (2018).

In this chapter, we consider three additional test procedures. One is, by considering the bivariate nature of interval-valued data, the Hotelling's  $T^2$  (HT) test. The other two are based on the univariate marginalization of interval-valued data, which we discussed in Chapter 2. For the marginalization, we use the histogrammed marginal. Thus the uniform kernel method (UK) and Gaussian kernel method (GK) by Jeon et al. (2015) are used to estimate the marginal distribution. We suggest using the Kolmogorov-Smirnov

(KS) distance between the kernel marginal distributions to test the equality of two populations, whose null distribution is approximated by a permutation procedure.

The remainder of this chapter is organized as follows. In Section 3.2, we introduce four methods to compare two interval-valued samples we consider in this paper. Two are direct applications of the existing methods and the other two are newly suggested based on univariate marginalization methods. In Section 3.3, we numerically compare the performance of four methods in various settings. In Section 3.4, we apply the methods to the BP data of female students in the US. In Section 3.5, we conclude the chapter with a summary.

## 3.2. Methods

We describe four methods for two-sample interval-valued data. The four methods are the CB, HT, UK, and GK tests. In the CB and HT, to remove the constraints in the variables, we transform interval-valued data  $(L, U]$  with  $L < U$  into  $(C, R)$  or  $(C, \log R)$  where  $C = (L + U)/2$  and  $R = (U - L)/2$ .

### 3.2.1. Combined (CB) test

Grzegorzewski (2018) propose a KS goodness-of-fit test for interval-valued data. This method applies the usual KS test to the center and half-range and combine the results.

Let us consider two independent samples of random intervals:  $\mathbf{X}_1, \dots, \mathbf{X}_m$  and  $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ . Let  $F_C$  and  $F_R$  be the cumulative distribution function (c.d.f.) of the center and half-range of  $\mathbf{X}$ ,

respectively. We define  $G_C$  and  $G_R$  similarly for  $\mathbf{Y}$ . Grzegorzewski (2018) suggests verifying the equivalence of the two populations by testing

$$\mathcal{H}_0 : F_C = G_C \text{ and } F_R = G_R.$$

The KS statistics for each hypothesis  $\mathcal{H}_{01} : F_C = G_C$  and  $\mathcal{H}_{02} : F_R = G_R$  are

$$\begin{aligned} T_1 &= D_{m,n}(\widehat{F}_{m,C}, \widehat{G}_{n,C}) = \left( \frac{mn}{m+n} \right)^{1/2} \sup_{t \in \mathcal{R}} |\widehat{F}_{m,C}(t) - \widehat{G}_{n,C}(t)|, \\ T_2 &= D_{m,n}(\widehat{F}_{m,R}, \widehat{G}_{n,R}) = \left( \frac{mn}{m+n} \right)^{1/2} \sup_{t \in \mathcal{R}} |\widehat{F}_{m,R}(t) - \widehat{G}_{n,R}(t)|, \end{aligned}$$

where  $\widehat{F}_{m,C}(t) = (1/m) \sum_{i=1}^m \mathbf{I}(C_{\mathbf{x}_i} \leq t)$ ,  $\widehat{F}_{m,R}(t) = (1/m) \sum_{i=1}^m \mathbf{I}(R_{\mathbf{x}_i} \leq t)$ ,  $\widehat{G}_{n,C}(t) = (1/n) \sum_{j=1}^n \mathbf{I}(C_{\mathbf{y}_j} \leq t)$ , and  $\widehat{G}_{n,R}(t) = (1/n) \sum_{j=1}^n \mathbf{I}(R_{\mathbf{y}_j} \leq t)$ .  $C_{\mathbf{x}_i}$  and  $R_{\mathbf{x}_i}$  represent the center and half-range of the interval  $\mathbf{X}_i$ , respectively, and  $C_{\mathbf{y}_j}$  and  $R_{\mathbf{y}_j}$  are similarly defined for the interval  $\mathbf{Y}_j$ . The asymptotic null distribution of  $T_1$  (or  $T_2$ ) under the null hypothesis is known as *Kolmogorov-Smirnov distribution* (Feller, 1948), where, for every fixed  $z \geq 0$ ,

$$\mathbf{P} \{T_1 \leq z\} \rightarrow L(z) = 1 - 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 z}, \quad (3.1)$$

as  $m \rightarrow \infty$ ,  $n \rightarrow \infty$  so that  $m/n \rightarrow a \in (0, \infty)$ . In the numerical study and data example followed, we use the permutation method to estimate the distribution of the test statistic  $T_1$  (or  $T_2$ ) due to the finiteness of the sample sizes.

To test the joint hypothesis  $\mathcal{H}_0$ , Grzegorzewski (2018) uses the Bonferroni procedure to combine the test results of  $\mathcal{H}_{01}$  and  $\mathcal{H}_{02}$ . Let  $p_1$  and  $p_2$  be the p-values related to  $T_1$  and  $T_2$ , respectively. Then, the overall p-value is set as  $p = 2 \min(p_1, p_2)$  to make the



overall size of the test be  $\alpha$ , taking the Bonferroni correction into account.  $\alpha$  is the assumed significant level, and we reject  $\mathcal{H}_0$  if  $p$  is small enough, such as  $p < \alpha$ .

### 3.2.2. Hotelling's $T^2$ (HT) test

Two-sample HT test is one of the most popular procedures to test the equality of two mean vectors of the populations. Here, we apply the HT to testing the equality of mean vectors of the center and log-transformed half-range of interval-valued data, which is a two-dimensional problem.

For notational simplicity, we abuse notations a little bit. Let  $\mathbf{X}_1, \dots, \mathbf{X}_m$ , and  $\mathbf{Y}_1, \dots, \mathbf{Y}_n$  be two samples of random intervals, where  $\mathbf{X}_i = (C_{1i}, \log R_{1i})$  and  $\mathbf{Y}_j = (C_{2j}, \log R_{2j})$ . We assume that the random intervals  $\mathbf{X}_1, \dots, \mathbf{X}_m$  ( $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ , respectively) are independently from the population with  $N_2(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})$  ( $N_2(\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}})$ , respectively), where  $N_2(\mu, \Sigma)$  denotes the bivariate normal distribution with mean vector  $\mu$  and covariance matrix  $\Sigma$ .

#### Equal covariance case

We assume that the covariances of the two populations are equal,  $\Sigma_{\mathbf{x}} = \Sigma_{\mathbf{y}}$ . Then the null hypothesis  $\mathcal{H}_0 : \mu_{\mathbf{x}} = \mu_{\mathbf{y}}$  can be testing using the  $HT_{\text{eq}}$ :

$$HT_{\text{eq}} = \frac{mn}{m+n} (\bar{\mathbf{X}} - \bar{\mathbf{Y}})^\top S_p^{-1} (\bar{\mathbf{X}} - \bar{\mathbf{Y}}),$$

where  $\bar{\mathbf{X}}$  and  $\bar{\mathbf{Y}}$  are the sample mean vectors of two samples, respectively, and  $S_p$  is the pooled covariance matrix calculated by

$$S_p = \frac{(m-1)S_{\mathbf{x}} + (n-1)S_{\mathbf{y}}}{m+n-2},$$

where  $S_{\mathbf{x}}$  and  $S_{\mathbf{y}}$  are the sample covariance matrices from  $\mathbf{X}_i$ s and  $\mathbf{Y}_j$ s, respectively.

Under the null hypothesis,

$$\frac{m+n-3}{m+n-2} HT_{\text{eq}} \sim F(2, m+n-3),$$

where  $F(2, m+n-3)$  is the F-distribution with parameters 2 and  $m+n-3$ .

### **Unequal covariance case**

If  $\Sigma_{\mathbf{x}} \neq \Sigma_{\mathbf{y}}$ , the HT statistic is computed as

$$HT_{\text{un}} = (\bar{\mathbf{X}} - \bar{\mathbf{Y}})^\top \left( \frac{S_{\mathbf{x}}}{m} + \frac{S_{\mathbf{y}}}{n} \right)^{-1} (\bar{\mathbf{X}} - \bar{\mathbf{Y}}).$$

Under the null hypothesis,

$$\frac{m+n-3}{2(m+n-2)} HT_{\text{un}} \sim F(2, \nu),$$

where  $\nu$  is an appropriately defined degrees of freedom.

### **3.2.3. Marginalization-based (UK and GK) test**

In this section, we propose a marginalization-based approach to test the equality of two interval-valued samples. As we discussed in Chapter 2, the marginalization means a univariate representation, say  $X$ , for an interval consisting of two variables,  $(L, U]$  (or  $(C, R)$ ). More specifically, we estimate the single-variable distribution function of  $X$ , denoted by  $F(x)$  for interval-valued data.

### **Two marginalizations**

We use two popular marginals: the uniform kernel estimator (UK) and Gaussian kernel estimator (GK). Both kernel estimators aim

to estimate the histogrammed marginalization, and a detailed description of both kernels is provided in Chapter 2.

The univariate marginal estimators of interval-valued data allow us to test the equality of distributions of two interval-valued samples. Let us consider two independent random intervals: first sample  $\mathbf{X}_1, \dots, \mathbf{X}_m$  is drawn from the population with c.d.f.  $F(\ell_1, u_2)$  where  $\ell_1$  and  $u_1$  indicate the lower and upper bound of the interval  $\mathbf{X}$ , respectively. The second sample  $\mathbf{Y}_1, \dots, \mathbf{Y}_n$  comes from the population with c.d.f.  $G(\ell_2, u_2)$  where  $\ell_2$  and  $u_2$  are defined similarly for  $\mathbf{Y}$ . We aim to test the null hypothesis  $\mathcal{H}_0 : F = G$ . To verify this, we suggest testing the equality of  $F_M(x)$  and  $G_M(y)$ , where  $F_M(x)$  and  $G_M(y)$  are the marginal distributions of  $F(\ell_1, u_1)$  and  $G(\ell_2, u_2)$ , respectively.

### Test statistic

The test statistic we propose is similar to the KS statistic:

$$T_M = D_{m,n}(\widehat{F}_{M,m}, \widehat{G}_{M,n}) = \left( \frac{mn}{m+n} \right)^{1/2} \sup_{t \in \mathcal{R}} |\widehat{F}_{M,m}(t) - \widehat{G}_{M,n}(t)|, \quad (3.2)$$

where  $\widehat{F}_{M,m}, \widehat{G}_{M,n}$  are the estimators of the marginal distributions  $F_M(x), G_M(y)$  based on  $\mathbf{X}_1, \dots, \mathbf{X}_m$ , and  $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ , respectively. For the GK, we are required to choose the bandwidth  $h$ . Here, we choose  $h$  to maximize the distance between the two estimated distributions. That is,  $h$  is selected as follows:

$$h_{\max} = \operatorname{argmax}_h \sup_{t \in \mathcal{R}} |\widehat{F}_{M,m}^h(t) - \widehat{G}_{M,n}^h(t)|,$$

where  $\widehat{F}_{M,m}^h$ ,  $\widehat{G}_{M,n}^h$  are the GK estimates with bandwidth  $h$ . Therefore, the test statistic  $T_M$  for the GK is as follows:

$$T_M = \left( \frac{mn}{m+n} \right)^{1/2} \sup_{t \in \mathcal{R}} |\widehat{F}_{M,m}^{h_{\max}}(t) - \widehat{G}_{M,n}^{h_{\max}}(t)|.$$

### Permutation procedure to approximate the null distribution

We use the permutation method to estimate the sampling distribution of the test statistic (3.2) under the null  $\mathcal{H}_0$ . The permutation procedure is straightforward, and the procedure is briefly described as follows. For the  $b$ -th permutation,  $b = 1, \dots, B$ , we combine all the  $m + n$  observations from both groups together, then take  $m$  random observations without replacement. This sample constitutes the first group and the remaining  $n$  observations are set as the second group. We compute the test statistic  $t_{M,b}$  (3.2) using these permuted samples and repeat this procedure  $B$  many times. The permutation distribution for the test statistic  $T_M$  under the null hypothesis  $\mathcal{H}_0$  is given by the empirical distribution of  $t_{M,1}, \dots, t_{M,B}$ . Now, let  $t_M^{\text{obs}}$  be the observed test statistic from the original two samples, the permutation p-value for hypothesis  $H_0$  is

$$p = \frac{\sum_{b=1}^B \mathbf{I}(t_{M,b} \geq t_M^{\text{obs}})}{B}.$$

In the numerical study, if we know the null distribution, the replication statistic  $t_{M,b}$  is computed by generating two new samples from the null itself rather than permuting two observed samples.

### 3.3. Numerical study

In this section, we compare the performance of four methods in the previous section under various situations. Two samples are generated according to the predetermined settings, and the size and power are evaluated. For the sample size  $(m, n)$ , we consider following 4 cases:  $(30, 30)$ ,  $(30, 120)$ ,  $(50, 50)$ ,  $(50, 200)$ .

For statistics,  $T_1$  ( $T_2$ ) and  $T_M$ , we numerically approximate its null distribution. We generate  $m$  and  $n$  samples under the null and calculate the test statistics,  $T_1$  ( $T_2$ ) and  $T_M$ . We repeat this procedure 20,000 many times to get the null distribution. For  $HT_{eq}$  ( $HT_{un}$ ), the simulated distribution is also used if the setting does not meet the underlying assumptions of the HT test.

The significance level  $\alpha$  is set as 5%. The size and power are evaluated as the rejection rate in 2,000 repetitions. All settings we consider for the study are summarized in Table 3.1, and their details and results are followed below.

#### 3.3.1. Normal distribution with equal covariances

We set a bivariate normal distribution for the center and log-transformed half-range. Then we compare the rejection power of four tests by varying the mean vector value of the second population, assuming that the covariances of the two populations are equal. By denoting the first population as  $\Pi_1$  and the second as  $\Pi_2$ , the setting is expressed as follows:

$$\Pi_1 : \begin{pmatrix} C_1 \\ \log R_1 \end{pmatrix} \sim N_2(\mu_1, \Sigma_1), \quad \Pi_2 : \begin{pmatrix} C_2 \\ \log R_2 \end{pmatrix} \sim N_2(\mu_2, \Sigma_2),$$

Table 3.1: Summary of the settings where  $\Sigma = (1 \quad -\rho; \rho \quad 1)$ . At the first column, the left character of the hyphen (-) denotes the distribution of  $(C, \log R)$  and the right represents the difference between the two populations. Among the left, N indicates “normal”, T for “T with df 5”, and SN for “skew-normal”. Among the right, C represents “mean of center”, R for “mean of range”, C.S for “mean and skewness of center”, COV for “covariance”, C.V for “mean and variance of center”, and R.V for “mean and variance of range”. The first population is denote by  $\Pi_1$  and the second is denoted by  $\Pi_2$  with  $\mu_1, \mu_2$  mean parameters,  $\Sigma_1, \Sigma_2$  covariance matrices, and  $\gamma_1, \gamma_2$  skewness parameters, respectively.

case	distribution of (C, log R)	$\Pi_1$			$\Pi_2$		
		$\mu_1$	$\Sigma_1$	$\gamma_1$	$\mu_2$	$\Sigma_2$	$\gamma_2$
(N-C)	Normal	(0, 0)	$\Sigma$	(0, 0)	( $\delta$ , 0)	$\Sigma$	(0, 0)
(N-R)	Normal	(0, 0)	$\Sigma$	(0, 0)	(0, $\delta$ )	$\Sigma$	(0, 0)
(T-C)	T with df 5	(0, 0)	$\Sigma$	(0, 0)	( $\delta$ , 0)	$\Sigma$	(0, 0)
(T-R)	T with df 5	(0, 0)	$\Sigma$	(0, 0)	(0, $\delta$ )	$\Sigma$	(0, 0)
(SN-C)	Skew normal	(0, 0)	$\Sigma$	(-0.6, -0.1)	( $\delta$ , 0)	$\Sigma$	(-0.6, -0.1)
(SN-C.S)	Skew normal	(0, 0)	$\Sigma$	(0, -0.1)	( $\delta$ , 0)	$\Sigma$	(-0.4 $\delta$ , -0.1)
(N-COV)	Normal	(0, 0)	$\Sigma$	(0, 0)	(0, 0)	$(1 + \delta)\Sigma$	(0, 0)
(N-C.V1)	Normal	(0, 0)	$\Sigma$	(0, 0)	( $\delta$ , 0)	$\begin{pmatrix} 1 + 2\delta & \sqrt{1 + 2\delta\rho} \\ \sqrt{1 + 2\delta\rho} & 1 \end{pmatrix}$	(0, 0)
(N-C.V2)	Normal	(0, 0)	$\begin{pmatrix} 4 & 2\rho \\ 2\rho & 1 \end{pmatrix}$	(0, 0)	( $\delta$ , 0)	$\begin{pmatrix} 4 - 2\delta & \sqrt{4 - 2\delta\rho} \\ \sqrt{4 - 2\delta\rho} & 1 \end{pmatrix}$	(0, 0)
(N-R.V)	Normal	(0, 0)	$\Sigma$	(0, 0)	(0, $\delta$ )	$\begin{pmatrix} 1 + 2\delta & \sqrt{1 + 2\delta\rho} \\ \sqrt{1 + 2\delta\rho} & 1 \end{pmatrix}$	(0, 0)

where mean and variance parameters are

$$\begin{aligned} \Pi_1 : \mu_1 &= (0, 0)^\top, \quad \Sigma_1 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \\ \text{(N-C)} \quad \Pi_2 : \mu_2 &= (\delta, 0)^\top, \quad \Sigma_2 = \Sigma_1 \\ \text{(N-R)} \quad \Pi_2 : \mu_2 &= (0, \delta)^\top, \quad \Sigma_2 = \Sigma_1. \end{aligned}$$

For  $\delta$ , following four values are used: (0, 0.5, 1.0, 1.5). The null hypothesis is naturally,  $\mathcal{H}_0 : \delta = 0$  for all four tests, the CB, HT, UK and GK tests. Thus when  $\delta = 0$ , we examine the size of

each test. That is, we check whether Type I error of each test is targeted at the chosen significance level  $\alpha$  or not. For  $\delta > 0$ , we assess the power of competing tests. To investigate the effect of the degree of correlation between the center and range, we use three values for  $\rho$ : (0, 0.4, 0.8).

Note that the mean vector in the second population ( $\Pi_2$ ) is set to either  $(\delta, 0)$  or  $(0, \delta)$ . The reason for varying mean parameter of center and half-range separately is that it affects the rejection power of each test differently, which is explained more below.

The results are shown in Table 3.2. Looking at the cases where  $\delta = 0$ , the size of each test is well controlled since the rejection rate is close to the significance level  $\alpha = 0.05$  in all cases. Under the other cases ( $\delta > 0$ ), it can be seen for every setting that the larger  $\delta$  is, the greater probability of rejection is. Similarly, a test becomes more powerful as more samples are available.

To summarize the winners based on the case where  $\rho = 0$ , the HT test shows the highest power among the four tests in both cases (N-C) and (N-R). This consequence is natural considering that other methods test the equality of distributions, while the HT test verifies only the equality of mean vectors between two populations and the data generation setting (a bivariate normal distribution with equal covariances) satisfies the underlying assumptions of the HT test. Note that in case (N-C), where two distributions differ in mean of the center, two marginal tests perform better than the CB, and are comparable to the HT. However, in case (N-R), where mean vectors are different at the range, the power of the CB is higher than that of the marginal tests.

Table 3.2: Simulation results. Power of each test in case of the bivariate normal distribution with equal covariances.

case	$(m, n)$	$\delta$	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(N-C)	(30, 30)	0.0	0.047	0.052	0.048	0.046	0.045	0.052	0.042	0.041	0.041	0.052	0.046	0.047
		0.5	0.293	0.381	<b>0.402</b>	0.387	0.296	<b>0.442</b>	0.420	0.412	0.242	<b>0.803</b>	0.520	0.498
		1.0	0.844	<b>0.931</b>	0.928	0.916	0.847	<b>0.966</b>	0.940	0.931	0.832	<b>1.000</b>	0.982	0.979
		1.5	0.997	<b>1.000</b>	0.998	0.998	0.997	<b>1.000</b>	0.999	0.999	0.995	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
	(30, 120)	0.0	0.044	0.053	0.046	0.043	0.052	0.053	0.048	0.045	0.048	0.053	0.047	0.043
		0.5	0.451	<b>0.577</b>	0.576	0.562	0.467	<b>0.659</b>	0.603	0.581	0.460	<b>0.958</b>	0.716	0.689
		1.0	0.974	<b>0.996</b>	0.993	0.992	0.974	<b>0.999</b>	0.995	0.993	0.977	<b>1.000</b>	0.999	0.999
		1.5	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
	(50, 50)	0.0	0.042	0.051	0.048	0.046	0.041	0.051	0.048	0.051	0.042	0.051	0.049	0.050
		0.5	0.466	<b>0.597</b>	0.594	0.575	0.454	<b>0.677</b>	0.626	0.606	0.446	<b>0.974</b>	0.752	0.745
		1.0	0.986	<b>0.996</b>	0.995	0.995	0.985	<b>1.000</b>	0.997	0.996	0.984	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
		1.5	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
	(50, 200)	0.0	0.040	0.047	0.047	0.048	0.046	0.047	0.046	0.047	0.040	0.047	0.053	0.051
		0.5	0.678	<b>0.810</b>	0.792	0.778	0.684	<b>0.888</b>	0.825	0.813	0.686	<b>1.000</b>	0.909	0.901
		1.0	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
		1.5	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
(N-R)	(30, 30)	0.0	0.047	0.052	0.048	0.046	0.045	0.052	0.042	0.041	0.041	0.052	0.046	0.047
		0.5	0.275	<b>0.365</b>	0.043	0.071	0.251	<b>0.430</b>	0.047	0.065	0.244	<b>0.795</b>	0.089	0.083
		1.0	0.840	<b>0.931</b>	0.109	0.315	0.840	<b>0.965</b>	0.159	0.347	0.840	<b>1.000</b>	0.409	0.433
		1.5	0.996	<b>0.999</b>	0.537	0.822	0.995	<b>1.000</b>	0.648	0.874	0.997	<b>1.000</b>	0.952	0.978
	(30, 120)	0.0	0.044	0.053	0.046	0.043	0.052	0.053	0.048	0.045	0.048	0.053	0.047	0.043
		0.5	0.470	<b>0.573</b>	0.060	0.123	0.462	<b>0.644</b>	0.075	0.146	0.465	<b>0.957</b>	0.155	0.217
		1.0	0.979	<b>0.995</b>	0.291	0.654	0.980	<b>0.998</b>	0.367	0.723	0.985	<b>1.000</b>	0.738	0.869
		1.5	<b>1.000</b>	<b>1.000</b>	0.924	0.986	<b>1.000</b>	<b>1.000</b>	0.965	0.992	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
	(50, 50)	0.0	0.042	0.051	0.048	0.046	0.041	0.051	0.048	0.051	0.042	0.051	0.049	0.050
		0.5	0.479	<b>0.602</b>	0.050	0.091	0.456	<b>0.680</b>	0.066	0.109	0.478	<b>0.971</b>	0.154	0.155
		1.0	0.981	<b>0.996</b>	0.281	0.607	0.983	<b>0.999</b>	0.394	0.680	0.987	<b>1.000</b>	0.752	0.791
		1.5	<b>1.000</b>	<b>1.000</b>	0.920	0.984	<b>1.000</b>	<b>1.000</b>	0.965	0.997	<b>1.000</b>	<b>1.000</b>	<b>0.999</b>	<b>1.000</b>
	(50, 200)	0.0	0.040	0.047	0.047	0.048	0.046	0.047	0.046	0.047	0.040	0.047	0.053	0.051
		0.5	0.694	<b>0.830</b>	0.093	0.201	0.699	<b>0.893</b>	0.111	0.238	0.713	<b>0.997</b>	0.264	0.334
		1.0	0.999	<b>1.000</b>	0.609	0.889	<b>1.000</b>	<b>1.000</b>	0.724	0.942	<b>1.000</b>	<b>1.000</b>	0.949	0.985
		1.5	<b>1.000</b>	<b>1.000</b>	0.999	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>

Looking closely at the properties of each test, in the CB and HT tests, the power in case (N-C) is almost the same to the power in (N-R) under the same simulation parameters. This result is also natural because both tests are designed with the same priority for the center and range. On the other hand, in the marginal tests, the power in case (N-C) is much higher than the power in case (N-R), especially when  $\delta$  is small. This implies that the two marginaliza-



tion methods, the UK and GK, are more sensitive to the change of the center rather than range. However, it is worth noting the performance of the marginal tests in case (N-R) with  $\rho = 0$ . That is, even if the range and center are independent, the power of the GK and UK is close to 1 as  $\delta$  grows. It should also be noted that the performance of the GK and UK is similar in case (N-C), but the GK performs much better than the UK in case (N-R).

Now, we examine the effect of correlation on the power of each test. It is found that the greater the correlation between the center and range, the higher the power of each test. This phenomenon can be explained using the Mahalanobis distance between the two mean vectors from  $\Pi_1$  and  $\Pi_2$ . In case (N-C), for instance, the distance is  $(\delta, 0) \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} (\delta, 0) = \delta^2 / (1 - \rho^2)$ , which increases as  $\rho$  gets larger. Specifically, when  $\rho$  is 0, 0.4, and 0.8, the corresponding distance is  $\delta^2$ ,  $1.2\delta^2$  and  $2.8\delta^2$ , respectively. Thus, it is evident to see that two population distributions are easily distinguished from each other, especially when  $\rho = 0.8$ . However, the effect of  $\rho$  in power differs for each test. The HT test shows the most significant increment in power among the four tests as  $\rho$  increases, which could be reasonable considering that the  $HT$  statistic is in the form of the Mahalanobis distance between two mean vectors. The followings are the UK and GK tests showing a similar increase. On the other hand, the power of the CB test hardly changes.

We hereafter would avoid discussion on the influence of different  $\rho$ s since results are almost the same in most of the following settings. Thus, the case of  $\rho = 0$  will be mainly discussed.

### 3.3.2. Non-normal cases

We examine the power of four tests in terms of tail thickness and skewness of an underlying bivariate distribution for the center and range.

#### Thickness of the tail

We use a bivariate t-distribution with the degrees of freedom 5 denoted by  $t_5$ , which has a thicker tail than the normal distribution. We assume two populations have equal covariance matrices. Other details regarding the setup are identical to the normal case. That is,

$$\Pi_1 : \begin{pmatrix} C_1 \\ \log R_1 \end{pmatrix} \sim t_5(\mu_1, \Sigma_1), \quad \Pi_2 : \begin{pmatrix} C_2 \\ \log R_2 \end{pmatrix} \sim t_5(\mu_2, \Sigma_2),$$

where mean and variance parameters are

$$\begin{aligned} \Pi_1 : \mu_1 &= (0, 0)^\top, \quad \Sigma_1 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \\ \text{(T-C)} \quad \Pi_2 : \mu_2 &= (\delta, 0)^\top, \quad \Sigma_2 = \Sigma_1 \\ \text{(T-R)} \quad \Pi_2 : \mu_2 &= (0, \delta)^\top, \quad \Sigma_2 = \Sigma_1. \end{aligned}$$

Since the Gaussian assumption is broken, the null distribution of  $HT_{\text{eq}}$  is calculated by the permutation method as mentioned earlier.

Table 3.3 represents the results. It is noticeable that the testing power decreases overall when compared to the case of the normal distribution.

Table 3.3: Simulation results. Power of each test in case of the bivariate t-distribution with df 5 with equal covariances

case	$(m, n)$	$\delta$	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(T-C)	(30, 30)	0.0	0.049	0.058	0.059	0.049	0.043	0.058	0.060	0.053	0.044	0.058	0.057	0.055
		0.5	0.262	0.253	<b>0.356</b>	0.256	0.214	0.304	<b>0.351</b>	0.298	0.245	<b>0.618</b>	0.428	0.398
		1.0	0.777	0.778	<b>0.870</b>	0.775	0.749	0.845	<b>0.886</b>	0.822	0.791	<b>0.993</b>	0.940	0.925
		1.5	0.983	0.977	<b>0.994</b>	0.979	0.977	0.991	<b>0.996</b>	0.988	0.983	<b>1.000</b>	0.998	0.998
	(30, 120)	0.0	0.042	0.053	0.047	0.053	0.044	0.053	0.054	0.055	0.041	0.053	0.048	0.050
		0.5	0.354	0.378	<b>0.486</b>	0.388	0.386	0.442	<b>0.507</b>	0.413	0.381	<b>0.807</b>	0.610	0.523
		1.0	0.937	0.924	<b>0.972</b>	0.947	0.950	0.956	<b>0.978</b>	0.961	0.951	<b>1.000</b>	0.995	0.988
		1.5	<b>1.000</b>	0.999	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
	(50, 50)	0.0	0.047	0.049	0.058	0.054	0.042	0.049	0.052	0.047	0.046	0.049	0.047	0.048
		0.5	0.385	0.396	<b>0.508</b>	0.393	0.390	0.460	<b>0.534</b>	0.444	0.388	<b>0.809</b>	0.641	0.603
		1.0	0.952	0.931	<b>0.976</b>	0.945	0.952	0.966	<b>0.985</b>	0.966	0.959	<b>1.000</b>	0.996	0.995
		1.5	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
	(50, 200)	0.0	0.043	0.043	0.047	0.050	0.044	0.043	0.050	0.057	0.040	0.043	0.052	0.048
		0.5	0.610	0.568	<b>0.713</b>	0.601	0.611	0.662	<b>0.736</b>	0.631	0.610	<b>0.958</b>	0.827	0.788
		1.0	0.997	0.997	<b>0.999</b>	0.996	0.995	<b>0.999</b>	<b>0.999</b>	0.997	0.998	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
		1.5	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
(T-R)	(30, 30)	0.0	0.049	0.058	0.059	0.049	0.043	0.058	0.060	0.053	0.044	0.058	0.057	0.055
		0.5	<b>0.257</b>	0.251	0.044	0.067	0.218	<b>0.291</b>	0.051	0.071	0.206	<b>0.611</b>	0.083	0.085
		1.0	<b>0.802</b>	0.777	0.119	0.214	0.763	<b>0.845</b>	0.152	0.225	0.773	<b>0.988</b>	0.329	0.316
		1.5	<b>0.989</b>	0.977	0.436	0.531	<b>0.986</b>	<b>0.986</b>	0.523	0.606	0.989	<b>1.000</b>	0.809	0.798
	(30, 120)	0.0	0.042	0.053	0.047	0.053	0.044	0.053	0.054	0.055	0.041	0.053	0.048	0.050
		0.5	<b>0.409</b>	0.388	0.058	0.100	0.392	<b>0.454</b>	0.076	0.118	0.396	<b>0.796</b>	0.133	0.174
		1.0	<b>0.959</b>	0.925	0.244	0.450	<b>0.957</b>	0.955	0.301	0.504	0.958	<b>0.999</b>	0.612	0.682
		1.5	<b>1.000</b>	0.998	0.804	0.830	<b>0.999</b>	<b>0.999</b>	0.870	0.894	<b>1.000</b>	<b>1.000</b>	0.980	0.976
	(50, 50)	0.0	0.047	0.049	0.058	0.054	0.042	0.049	0.052	0.047	0.046	0.049	0.047	0.048
		0.5	0.407	<b>0.418</b>	0.066	0.102	0.419	<b>0.481</b>	0.080	0.105	0.400	<b>0.830</b>	0.134	0.129
		1.0	<b>0.961</b>	0.941	0.250	0.369	0.963	<b>0.964</b>	0.321	0.410	0.961	<b>1.000</b>	0.594	0.564
		1.5	<b>1.000</b>	0.999	0.803	0.808	<b>1.000</b>	<b>1.000</b>	0.863	0.860	<b>1.000</b>	<b>1.000</b>	0.980	0.966
	(50, 200)	0.0	0.043	0.043	0.047	0.050	0.044	0.043	0.050	0.057	0.040	0.043	0.052	0.048
		0.5	<b>0.619</b>	0.570	0.087	0.158	0.605	<b>0.655</b>	0.102	0.180	0.610	<b>0.968</b>	0.208	0.246
		1.0	<b>0.998</b>	0.994	0.499	0.658	0.996	<b>0.997</b>	0.596	0.730	0.998	<b>1.000</b>	0.858	0.870
		1.5	<b>1.000</b>	<b>1.000</b>	0.981	0.965	<b>1.000</b>	<b>1.000</b>	0.995	0.979	<b>1.000</b>	<b>1.000</b>	0.999	<b>1.000</b>

Based on the case where  $\rho$  is 0, first, unlike the normal case where the HT test shows the highest power, the UK test outperforms the other three tests in case (T-C), where mean vectors are different at the center. In case (T-R), where mean vectors are different at the range, the CB test is most powerful. Performance degradation of the HT is obvious since the Gaussian assumption is not satisfied. Second, in case (T-C), the power of the UK test

uniformly dominates that of the GK, contrary to the similar performance of the UK and GK tests in the previous normal case (N-C). The lesser performance of the GK test relates to the dependency of the GK estimator on the Gaussian kernel. Besides, as in the previous results, in case (T-R), the performance of the marginal tests is much lower than that of the two other tests except the case with large  $\delta = 1.5$ .

Meanwhile, when  $\rho = 0.8$ , the power of the HT test is higher than that of the other tests. This is because as  $\rho$  gets larger, the increase of power of the HT is much more substantial than the other tests, as described above.

### Skewness

For the center and log-transformed half-range, we set the following bivariate skew-normal distribution. We use CP parameterization to fix the marginal parameters at prescribed values (Azzalini and Capitanio, 1999). That is,

$$\begin{pmatrix} C \\ \log R \end{pmatrix} \sim SN \left[ \mu = \begin{pmatrix} \mu_C \\ \mu_R \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \gamma = \begin{pmatrix} \gamma_C \\ \gamma_R \end{pmatrix} \right],$$

where  $(\gamma_C, \gamma_R)^\top$  represents skewness of the marginal distribution of the center and log-transformed half-range, respectively. For the sake of simplicity, we only consider two cases for sample size,  $m = 30, n = 30$  and  $m = 30, n = 120$ . Also, only the cases where mean vectors are different at the center are under consideration. We additionally consider the case where skewness and mean of the center are varying together, which reflects the feature of the real data example described in the next section.

(SN-C) Only mean of the center is different while covariance and skewness are the same in two populations:

$$\Pi_1 : \mu_1 = (0, 0)^\top, \quad \Sigma_1 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad \gamma_1 = (-0.6, -0.1)^\top$$

$$\Pi_2 : \mu_2 = (\delta, 0)^\top, \quad \Sigma_2 = \Sigma_1, \quad \gamma_2 = \gamma_1.$$

(SN-C.S) Skewness of the center as well as mean of the center are different in two populations, and two covariances are equal:

$$\Pi_1 : \mu_1 = (0, 0)^\top, \quad \Sigma_1 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad \gamma_1 = (0, -0.1)^\top$$

$$\Pi_2 : \mu_2 = (\delta, 0)^\top, \quad \Sigma_2 = \Sigma_1, \quad \gamma_2 = (-2\delta/5, -0.1)^\top.$$

Results are shown in Table 3.4. In case of (SN-C), where two populations is different at mean of the center, the result is similar

Table 3.4: Simulation results. Power of each test in case of the bivariate skew-normal distribution with equal covariances.

case	$(m, n)$	$\delta$	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(SN-C)	(30, 30)	0.0	0.045	0.047	0.042	0.040	0.044	0.051	0.045	0.046	0.040	0.047	0.057	0.053
		0.5	0.303	<b>0.387</b>	0.345	0.341	0.306	<b>0.419</b>	0.373	0.359	0.300	<b>0.801</b>	0.487	0.474
		1.0	0.889	<b>0.927</b>	0.908	0.907	0.890	<b>0.962</b>	0.924	0.916	0.894	<b>1.000</b>	0.972	0.965
		1.5	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
	(30, 120)	0.0	0.046	0.052	0.051	0.052	0.043	0.056	0.052	0.052	0.041	0.053	0.049	0.049
		0.5	0.498	<b>0.566</b>	0.515	0.523	0.498	<b>0.648</b>	0.556	0.529	0.499	<b>0.958</b>	0.669	0.634
		1.0	0.993	<b>0.999</b>	0.993	0.996	0.993	<b>0.999</b>	0.992	0.991	0.994	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
		1.5	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
(SN-C.S)	(30, 30)	0.0	0.045	0.050	0.044	0.043	0.042	0.052	0.044	0.049	0.041	0.050	0.049	0.052
		0.5	0.317	0.366	<b>0.424</b>	0.409	0.315	0.423	<b>0.437</b>	0.434	0.310	<b>0.803</b>	0.536	0.522
		1.0	0.886	0.922	<b>0.937</b>	0.923	0.888	<b>0.962</b>	0.952	0.944	0.891	<b>1.000</b>	0.983	0.982
		1.5	0.998	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	0.998	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	0.999	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
	(30, 120)	0.0	0.048	0.051	0.058	0.056	0.046	0.051	0.056	0.055	0.043	0.056	0.046	0.046
		0.5	0.532	0.582	<b>0.634</b>	0.625	0.529	<b>0.656</b>	0.652	0.629	0.531	<b>0.962</b>	0.764	0.731
		1.0	0.990	<b>0.995</b>	<b>0.995</b>	<b>0.995</b>	0.990	<b>0.999</b>	0.997	0.996	0.992	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
		1.5	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>

to the normal case where the HT test shows the best performance and the power of two marginal tests is better than that of the CB.

Next, look at the case (SN-C.S), where two populations differ in skewness of the center as well as mean of the center. Skewness of the center of the second population is set to gradually increase to the left, referring to the data example in the next section. When  $\rho = 0$ , we find that the UK and GK tests, which are marginal tests, are superior to the other two tests, unlike the previous case (SN-C). As in the previous cases, when  $\rho = 0.8$ , the power of the HT test is the highest.

### 3.3.3. Normal distribution with unequal covariances

We also set a bivariate normal distribution for the center and log-transformed half-range, but this time we assume that covariances of two populations are not equal. We consider the following four cases, including the case that represents the characteristics of the real data example in the next section, such as the above case (SN-C.S). We use only two cases for sample size for simplicity:  $m = 30, n = 30$  and  $m = 30, n = 120$ .

(N-COV) The covariance matrices are unequal while the mean vectors are equal:

$$\begin{aligned} \Pi_1 : \mu_1 &= (0, 0)^\top, \quad \Sigma_1 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \\ \Pi_2 : \mu_2 &= (0, 0)^\top, \quad \Sigma_2 = (1 + \delta)\Sigma_1. \end{aligned}$$

(N-C.V1) The mean and variance of the center are different in two populations. In the second population, both the mean and

variance of the center increase:

$$\begin{aligned}\Pi_1 : \mu_1 &= (0, 0)^\top, \quad \Sigma_1 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \\ \Pi_2 : \mu_2 &= (\delta, 0)^\top, \quad \Sigma_2 = \begin{pmatrix} 1 + 2\delta & \sqrt{1 + 2\delta}\rho \\ \sqrt{1 + 2\delta}\rho & 1 \end{pmatrix}.\end{aligned}$$

(N-C.V2) In the second population, the mean of center increases while the variance of center decreases:

$$\begin{aligned}\Pi_1 : (\mu_C, \mu_R)^\top &= (0, 0)^\top, \quad \Sigma_1 = \begin{pmatrix} 4 & 2\rho \\ 2\rho & 1 \end{pmatrix} \\ \Pi_2 : (\mu_C, \mu_R)^\top &= (\delta, 0)^\top, \quad \Sigma_2 = \begin{pmatrix} 4 - 2\delta & \sqrt{4 - 2\delta}\rho \\ \sqrt{4 - 2\delta}\rho & 1 \end{pmatrix}.\end{aligned}$$

(N-R.V) The mean and variance of the range differ in two populations. In the second population, both the mean and variance of the range increase:

$$\begin{aligned}\Pi_1 : \mu_1 &= (0, 0)^\top, \quad \Sigma_1 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \\ \Pi_2 : \mu_2 &= (0, \delta)^\top, \quad \Sigma_2 = \begin{pmatrix} 1 & \sqrt{1 + 2\delta}\rho \\ \sqrt{1 + 2\delta}\rho & 1 + 2\delta \end{pmatrix}.\end{aligned}$$

Table 3.5 shows the results. As mentioned earlier, we mainly describe the results based on the cases where  $\rho = 0$ . The most interesting result is the case (N-COV), where only the covariances differ in two populations. In case (N-COV), the power of the marginal tests is much higher than that of other tests. Among the marginal tests, the GK outperforms the UK. This result means that the

Table 3.5: Simulation results. Power of each test in case of the bivariate normal distribution with unequal covariances.

case	$(m, n)$	$\delta$	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
			CB	HT	UK	GK	CB	HT	UK	GK	CB	HT	UK	GK
(N-COV)	(30, 30)	0.0	0.047	0.052	0.048	0.046	0.045	0.052	0.042	0.041	0.041	0.052	0.046	0.047
		0.5	0.061	0.050	0.080	<b>0.104</b>	0.053	0.050	0.082	<b>0.112</b>	0.047	0.050	0.098	<b>0.115</b>
		1.0	0.077	0.046	0.155	<b>0.227</b>	0.086	0.046	0.157	<b>0.256</b>	0.071	0.046	0.217	<b>0.280</b>
		1.5	0.122	0.052	0.239	<b>0.408</b>	0.115	0.052	0.263	<b>0.431</b>	0.097	0.052	0.384	<b>0.496</b>
	(30, 120)	0.0	0.044	0.054	0.046	0.043	0.052	0.054	0.048	0.045	0.048	0.054	0.047	0.043
		0.5	0.051	0.045	0.072	<b>0.117</b>	0.047	0.045	0.078	<b>0.142</b>	0.040	0.045	0.111	<b>0.178</b>
		1.0	0.092	0.047	0.161	<b>0.325</b>	0.085	0.047	0.176	<b>0.407</b>	0.077	0.047	0.284	<b>0.483</b>
		1.5	0.138	0.044	0.306	<b>0.592</b>	0.144	0.044	0.345	<b>0.689</b>	0.127	0.044	0.546	<b>0.784</b>
(N-C.V1)	(30, 30)	0.0	0.047	0.052	0.048	0.046	0.045	0.052	0.042	0.041	0.041	0.052	0.046	0.047
		0.5	0.262	0.258	<b>0.420</b>	0.401	0.265	0.305	<b>0.385</b>	0.343	0.241	<b>0.601</b>	0.402	0.374
		1.0	0.700	0.664	<b>0.865</b>	0.837	0.716	0.734	<b>0.834</b>	0.818	0.715	<b>0.971</b>	0.865	0.841
		1.5	0.944	0.897	<b>0.988</b>	0.981	0.946	0.938	<b>0.982</b>	0.977	0.938	<b>1.000</b>	0.990	0.987
	(30, 120)	0.0	0.044	0.054	0.046	0.043	0.052	0.054	0.048	0.045	0.048	0.054	0.047	0.043
		0.5	0.436	0.489	<b>0.611</b>	0.594	0.448	0.553	<b>0.564</b>	0.536	0.450	<b>0.899</b>	0.573	0.556
		1.0	0.949	0.956	<b>0.986</b>	<b>0.986</b>	0.951	<b>0.984</b>	0.979	0.975	0.946	<b>1.000</b>	0.981	0.980
		1.5	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	0.999	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
(N-C.V2)	(30, 30)	0.0	0.046	0.052	0.043	0.045	0.041	0.052	0.041	0.046	0.040	0.052	0.049	0.050
		0.5	0.098	0.142	<b>0.152</b>	0.149	0.105	0.154	<b>0.174</b>	0.161	0.098	<b>0.298</b>	0.202	0.198
		1.0	0.438	0.470	<b>0.557</b>	0.528	0.430	0.551	<b>0.613</b>	0.610	0.430	<b>0.896</b>	0.710	0.719
		1.5	0.934	0.898	<b>0.956</b>	0.948	0.938	0.939	0.977	<b>0.978</b>	0.935	<b>0.999</b>	0.996	0.997
	(30, 120)	0.0	0.043	0.054	0.045	0.047	0.048	0.054	0.045	0.048	0.042	0.054	0.045	0.049
		0.5	0.180	0.185	<b>0.256</b>	0.243	0.181	0.209	<b>0.268</b>	0.253	0.157	<b>0.422</b>	0.299	0.281
		1.0	0.673	0.612	<b>0.759</b>	0.745	0.676	0.687	<b>0.802</b>	0.784	0.666	<b>0.968</b>	0.870	0.854
		1.5	0.989	0.940	<b>0.994</b>	0.993	0.992	0.973	<b>0.998</b>	<b>0.998</b>	0.992	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
(N-R.V)	(30, 30)	0.0	0.047	0.052	0.048	0.046	0.045	0.052	0.042	0.041	0.041	0.052	0.046	0.047
		0.5	0.285	0.264	0.067	<b>0.351</b>	0.271	0.303	0.084	<b>0.334</b>	0.260	<b>0.597</b>	0.166	0.256
		1.0	0.732	0.653	0.195	<b>0.886</b>	0.721	0.736	0.265	<b>0.894</b>	0.724	<b>0.969</b>	0.552	0.859
		1.5	0.947	0.907	0.517	<b>0.993</b>	0.950	0.938	0.635	<b>0.994</b>	0.948	<b>1.000</b>	0.905	0.995
	(30, 120)	0.0	0.044	0.054	0.046	0.043	0.052	0.054	0.048	0.045	0.048	0.054	0.047	0.043
		0.5	0.466	0.492	0.079	<b>0.614</b>	0.473	0.558	0.120	<b>0.620</b>	0.458	<b>0.892</b>	0.250	0.602
		1.0	0.963	0.958	0.425	<b>0.994</b>	0.962	0.979	0.535	<b>0.992</b>	0.962	<b>0.998</b>	0.852	0.994
		1.5	<b>1.000</b>	<b>1.000</b>	0.932	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	0.965	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	0.999	<b>1.000</b>

marginal tests, especially the GK test, effectively detect the difference in covariance over the other tests. On the contrary, the HT test, which tests the difference between two mean vectors, is incapable of detecting covariance differences between two populations, as it shows the power same to the size. In cases of (N-C.V1) and (N-C.V2), where variance of the center in the second population also varies (increases or decreases) with the mean change,



the power of the marginal tests is higher than that of the other tests. Recall that the power of the HT test is the highest in case (N-C), where only the mean of the center varies. Finally, in case (N-R.V), where both the mean and variance of the range vary, the GK test shows much higher power than the other tests, unlike the low performance in case (N-R), where only the mean of the range varies. When  $\rho = 0.8$ , the HT test has the highest power in all other cases except the case (N-COV), where there is no difference in the two mean vectors.

To summarize the numerical study, Table 3.6 shows the best and worst performers in each case of the numerical study. The marginal tests, the UK and GK tests, show good performance when compared to the existing methods. In particular, the power of the marginal tests is higher than that of the other methods if two distributions differ by more than one factor: mean, covariance, and skewness, etc. Note also that the marginal tests detect the center difference better than the range difference.

Table 3.6: Summary of the results. The best and worst tests are represented for each case. At the second column, the left character of the hyphen (-) denotes the distribution of  $(C, \log R)$  and the right represents the difference between the two populations.

	case	$\rho = 0$		$\rho = 0.4$		$\rho = 0.8$	
		best	worst	best	worst	best	worst
equal covariances	(N-C)	HT ( $\approx$ UK, GK)	CB	HT	CB	HT	CB
	(N-R)	HT	UK	HT	UK	HT	UK
	(T-C)	<b>UK</b>	HT( $\approx$ CB)	<b>UK</b>	CB	HT	CB
	(T-R)	CB( $\approx$ HT)	UK	HT( $\approx$ CB)	UK	HT	UK( $\approx$ GK)
	(SN-C)	HT	CB	HT	CB	HT	CB
	(SN-C.S)	<b>UK</b> ( $\approx$ <b>GK</b> )	CB	HT( $\approx$ UK)	CB	HT	CB
unequal covariances	(N-COV)	<b>GK</b>	HT	<b>GK</b>	HT	<b>GK</b>	HT
	(N-C.V1)	<b>UK</b> ( $\approx$ <b>GK</b> )	CB	<b>UK</b>	CB	HT	CB
	(N-C.V2)	<b>UK</b> ( $\approx$ <b>GK</b> )	CB	<b>UK</b> ( $\approx$ <b>GK</b> )	CB	HT	CB
	(N-R.V)	<b>GK</b>	UK	<b>GK</b>	UK	HT	UK

### 3.4. Data example

We apply the four two-sample comparison methods discussed above to analyze a real dataset. We also use the BP data described in Chapter 2. Thus, the description of the BP data is replaced with the contents of Chapter 2. Recall that through the descriptive statistics analysis in the previous chapter, we found that mean of African-Americans is higher than that of Caucasian.

Table 3.7 shows the results when two-sample comparison methods are applied to the BP data. In all tests, the p-values are much smaller than 0.001, confirming the significant difference between the two groups.

Table 3.7: Two-sample tests for the whole BP data.

	CB	HT	UK	GK
p-value	< 0.001	< 0.001	< 0.001	< 0.001

### 3.4.1. Sub-sampling

Since the sizes of the two samples ( $m = 1,112$ ,  $n = 1,144$ ) are very large compared to the typical sample size, the p-value of each test is so underestimated that it is difficult to compare the performance of four tests. Thus we sub-sample  $m'$  and  $n'$  from the two original samples and perform four tests based on the sub-sampled observations and compare the rejection power of each test again.

Now the original sample is considered as a population. Its descriptive statistics are given in Table 3.8 and contour plots are displayed in Figure 3.1, which are the summary of the data transformed to  $(C, \log R)$ .

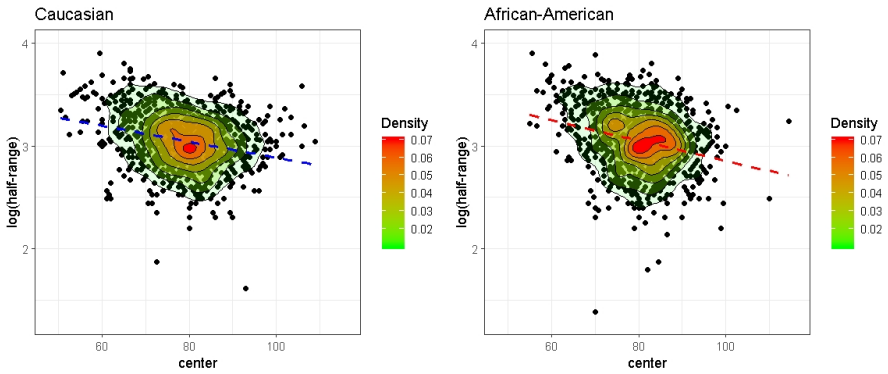


Figure 3.1: Contour plots of the two groups of BP data

Table 3.8: Descriptive statistics of  $(C, \log R)$  for the BP data by race, where  $C$  is the center and  $R$  is the half-range. This is a summary of the population of the sub-samples. The correlation coefficient is for the center and log-transformed half-range.

	Caucasian ( $m = 1,112$ )	African-American ( $n = 1,144$ )
mean	(78.67, 3.05)	(80.13, 3.05)
covariance	$\begin{pmatrix} 82.75 & -0.64 \\ -0.64 & 0.07 \end{pmatrix}$	$\begin{pmatrix} 64.56 & -0.64 \\ -0.64 & 0.09 \end{pmatrix}$
correlation coefficient	-0.26	-0.27
skewness	(-0.07, -0.27)	(-0.16, -0.35)

Looking into the details of the summary, mean of the center of African-American girls is larger than that of Caucasians while variance of the center of African-American girls is less than that of Caucasians. Mean of the log-transformed half-range is the same in both groups, and the distributions of the center and log-transformed half-range of African-American are skewed to the left than those of Caucasians. Correlation coefficients between the center and log-transformed half-range for the two groups are as low as -0.26 and -0.27, respectively.

Table 3.9 summarizes the rejection power depending on the size of sub-samples among 2,000 replicates. The two marginal tests, the UK and GK, which show similar power, perform best, followed by the HT and CB tests. This result is consistent with the previous numerical study, especially the cases (SN-C.S) and (N-C.V2) with small  $\rho$ . Recall that in case (SN-C.S),  $(C, \log R)$  has a skew-normal distribution and mean and skewness of the center differ in two

Table 3.9: Power of four two-sample testing methods for different sub-sample sizes.

$m'$	$n'$	CB	HT	UK	GK
30	30	0.074	0.082	0.097	<b>0.098</b>
30	120	0.084	0.116	0.130	<b>0.132</b>
50	50	0.100	0.110	0.143	<b>0.150</b>
50	200	0.125	0.162	<b>0.194</b>	0.188
100	100	0.146	0.176	0.232	<b>0.238</b>
100	400	0.215	0.264	0.320	<b>0.323</b>
300	300	0.404	0.436	<b>0.593</b>	0.554

populations. In case (N-C.V2)  $(C, \log R)$  is normally distributed and mean and variance of the center differ in two populations (In the second population, mean of the center increases while variance of the center decreases).

### 3.5. Conclusion

In this chapter, we propose a marginalization-based test to verify whether two samples of interval-valued data come from the same distribution. Also, we apply the Hotelling's  $T^2$  test to examine the equality of mean vectors of the center and range of interval-valued data.

Numerical study and real data analysis show that the marginal tests perform better than the existing methods, especially when two population distributions are different due to more than one factor, such as mean, covariance, skewness, and so on. This implies that the marginal tests can be more suitable for testing real prob-

lems of interval-valued data. Among the marginal tests, the GK test, which selects the best bandwidth  $h$  per test, shows greater flexibility to detect the difference between two distributions than the UK.

## Chapter 4

# Testing for Stochastic Order in Interval-Valued Data

### 4.1. Introduction

In this chapter, we discuss the two-sample order tests for interval-valued data. It is a fundamental problem in statistics to test the stochastic order of two populations as well as to verify the equality of the two. However, there is little research for interval-valued data.

In this chapter, we propose a method to test the stochastic order of two samples of interval-valued data. The remainder of the chapter is organized as follows. In section 4.2, we define the stochastic order of interval-valued data. In section 4.3, we propose a statistic to test the order of interval-valued data and derive its asymptotic null distribution using the theory on U-statistic. In

section 4.4, we examine the performance of the order test through a numerical study. In section 4.5, we apply the method to the BP data used in the previous chapters. In section 4.6, we conclude the paper with a summary.

## 4.2. Simple stochastic order

Before we introduce the notion of the stochastic order for interval-valued data, we look at the stochastic order for the usual univariate case. Let  $X$  and  $Y$  be two univariate random variables such that

$$Pr(X > z) \leq Pr(Y > z) \quad \text{for all } z \in \mathcal{R}.$$

Then  $Y$  is said to be *stochastically greater than*  $X$  (denoted by  $X \leq_{st} Y$ ). If additionally  $Pr(X > z) < Pr(Y > z)$  for some  $z$ , then  $Y$  is said to be *stochastically strictly greater than*  $X$ .

The stochastic order for interval-valued data can be defined similarly. Let  $\mathbf{x} = (\ell_1, u_1]$  and  $\mathbf{y} = (\ell_2, u_2]$  be two intervals. Then we denote  $\mathbf{x} < \mathbf{y}$  and say  $\mathbf{y}$  is *greater* than  $\mathbf{x}$  if  $\ell_1 < \ell_2$  and  $u_1 < u_2$ . Now, let  $\mathbf{X}$  and  $\mathbf{Y}$  be two random intervals such that

$$Pr(\mathbf{X} > \mathbf{z}) \leq Pr(\mathbf{Y} > \mathbf{z}) \quad \text{for all interval } \mathbf{z}. \quad (4.1)$$

Then  $\mathbf{Y}$  is said to be *stochastically greater than*  $\mathbf{X}$  and denoted by  $\mathbf{X} \leq_{st} \mathbf{Y}$ . Let  $\bar{F}(\mathbf{x})$  and  $\bar{G}(\mathbf{y})$  be the survival functions of the random intervals  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively, then (4.1) is the same as

$$\bar{F}(\ell, u) \leq \bar{G}(\ell, u) \quad \text{for all } (\ell, u). \quad (4.2)$$

We can illustrate the order of the intervals as follows (see Figure 4.1). Let the interval  $(\ell_1, u_1]$  denoted by  $(\ell_1, u_1)$  in the plane.



Note that in the plane, interval-valued data is displayed at the top of the  $u = \ell$  line due to  $\ell < u$  constraint. All interval-valued data of the half-plane can be divided into the following three types according to the order relation with the interval  $I_1 = (\ell_1, u_1)$ .

- The intervals in region A are greater than the interval  $I_1 = (\ell_1, u_1)$ .
- The intervals in region C are less than the interval  $I_1 = (\ell_1, u_1)$ .
- The intervals in region B or D do not have an order relation with the interval  $I_1 = (\ell_1, u_1)$ : The intervals in region B contain the interval  $I_1 = (\ell_1, u_1)$ , while the intervals in region D are included in the interval  $I_1 = (\ell_1, u_1)$ .

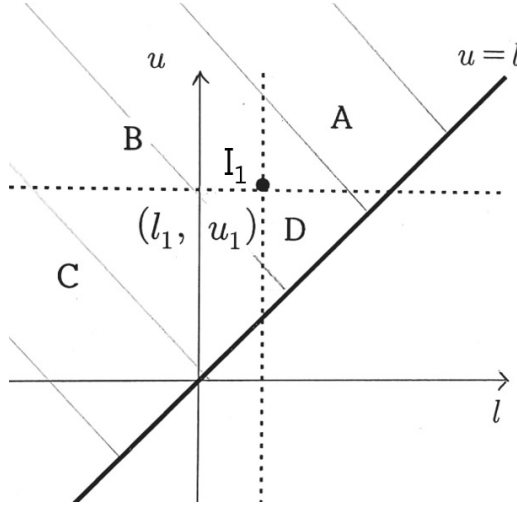


Figure 4.1: A graphical illustration of the order of interval-valued data

### 4.3. Test statistic

Let us consider two independent samples of random intervals. Suppose that a first sample  $\mathbf{X}_i = (L_{1i}, U_{1i}]$ ,  $i = 1, \dots, m$  has distribution with c.d.f.  $F(\mathbf{x})$  and the second sample  $\mathbf{Y}_j = (L_{2j}, U_{2j}]$ ,  $j = 1, \dots, n$  has distribution with c.d.f.  $G(\mathbf{y})$ . We want to verify the null hypothesis that both samples come from the same distribution,  $\mathcal{H}_0 : F(\mathbf{z}) = G(\mathbf{z})$  for all  $\mathbf{z}$  versus the alternative hypothesis that  $\mathbf{Y}$  is stochastically (strictly) greater than  $\mathbf{X}$ ,  $\mathcal{H}_1 : \overline{F}(\mathbf{z}) < \overline{G}(\mathbf{z})$  for some interval  $\mathbf{z}$ .

Let  $(\ell_{1i}, u_{1i}]$ ,  $i = 1, \dots, m$  and  $(\ell_{2j}, u_{2j}]$ ,  $j = 1, \dots, n$  be the observed intervals for two samples. The statistic we propose to test the stochastic order is

$$T = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n S_{ij}, \quad (4.3)$$

where

$$S_{ij} = \begin{cases} 1 & \text{if } \ell_{1i} < \ell_{2j} \text{ and } u_{1i} < u_{2j}, \\ -1 & \text{if } \ell_{1i} > \ell_{2j} \text{ and } u_{1i} > u_{2j}, \\ 0 & \text{otherwise.} \end{cases}$$

Note that, when  $F = G$ ,  $Pr(S_{ij} = 1) = Pr(S_{ij} = -1)$  and  $E(T) = 0$ .

The asymptotic null distribution of the statistic  $T$  is the normal distribution with mean 0 and variance given below. The statistic  $T$  is a U-statistic and it is proven based on the asymptotic of the U-statistic. We refer the results reported in Chapter 6 of Lehmann (1999). Suppose  $X_1, \dots, X_m$  are independently and identically distributed (i.i.d.) from c.d.f.  $F(x)$  and  $Y_1, \dots, Y_n$  are i.i.d. from  $G(y)$ . Let  $\phi(x_1, \dots, x_a; y_1, \dots, y_b)$  be a kernel of  $a + b$  arguments, which

is symmetric in each of the two groups of arguments such that

$$\theta = \theta(F, G) = E[\phi(X_1, \dots, X_a; Y_1, \dots, Y_b)].$$

and define its U-statistic

$$U_{m,n} = \binom{m}{a}^{-1} \binom{n}{b}^{-1} \sum_{C_{m,a}} \sum_{C_{n,b}} \phi(X_{i_1}, \dots, X_{i_a}; Y_{j_1}, \dots, Y_{j_b}), \quad (4.4)$$

which is an unbiased estimator of  $\theta$ . The variance of the U-statistic (4.4) is

$$Var(U_{m,n}) = \sum_{i=1}^a \sum_{j=1}^b \frac{\binom{a}{i} \binom{m-a}{a-i}}{\binom{m}{a}} \frac{\binom{b}{j} \binom{n-b}{b-j}}{\binom{n}{b}} \sigma_{ij}^2,$$

where

$$\sigma_{ij}^2 = Cov[\phi(X_1, \dots, X_i, X_{i+1}, \dots, X_a; Y_1, \dots, Y_j, Y_{j+1}, \dots, Y_b), \\ \phi(X_1, \dots, X_i, X'_{i+1}, \dots, X'_a; Y_1, \dots, Y_j, Y'_{j+1}, \dots, Y'_b)],$$

where  $X'_i$ 's and  $Y'_j$ 's are independent copies of  $X_i$ 's and  $Y_j$ 's. The theorem below from Chapter 6 of Lehmann (1999) finds the asymptotic distribution of the U-statistic in (4.4) above.

**Theorem 4.1.**  $\sqrt{N}(U_{m,n} - \theta)$  converges in distribution to the normal distribution with mean 0 and variance  $\sigma^2 = \frac{a^2}{\rho} \sigma_{10}^2 + \frac{b^2}{1-\rho} \sigma_{01}^2$ , where

$$\sigma_{10}^2 = Cov[\phi(X_1, X_2, \dots, X_a; Y_1, \dots, Y_b), \phi(X_1, X'_2, \dots, X'_a; Y'_1, \dots, Y'_b)] \in (0, \infty), \\ \sigma_{01}^2 = Cov[\phi(X_1, \dots, X_a; Y_1, Y_2, \dots, Y_b), \phi(X'_1, \dots, X'_a; Y_1, Y'_2, \dots, Y'_b)] \in (0, \infty)$$

and  $\sigma_{ab}^2 < \infty$  as  $m/N \rightarrow \rho \in (0, 1)$  and  $N = (m + n) \rightarrow \infty$ .

Applying the theorem above to our case, we can compute the asymptotic null distribution of our  $T$  statistic.

**Theorem 4.2.** Under the null hypothesis that  $\mathcal{H}_0 : F = G$ , if  $m/N \rightarrow \rho \in (0, 1)$  as  $N = (m + n) \rightarrow \infty$ , then

$$\sqrt{N}T \xrightarrow{d} N\left(0, f(p)\frac{N^2}{mn}\right), \quad (4.5)$$

where  $f(p) = (8p^3/3 + 4p^2(1 - 2p))$  and  $p = \Pr(S_{ij} = 1) = \Pr(S_{ij} = -1)$ .

*Proof.* If we let  $\phi(\mathbf{x} = (\ell_1, u_1]; \mathbf{y} = (\ell_2, u_2]) = \mathbf{I}(\ell_1 < \ell_2, u_1 < u_2) - \mathbf{I}(\ell_1 > \ell_2, u_1 > u_2)$ , the two sample U-statistic

$$U_{m,n} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n \phi(\mathbf{X}_i; \mathbf{Y}_j), \quad (4.6)$$

with  $a = b = 1$  equals to  $T$  in (4.3). By applying Theorem 4.1, we have

$$\sqrt{N}(U_{m,n} - \theta) \xrightarrow{d} N\left(0, \frac{\sigma_{10}^2}{\rho} + \frac{\sigma_{01}^2}{1 - \rho}\right),$$

where  $\theta = E(\phi(\mathbf{X}; \mathbf{Y}))$ ,  $\rho = \lim \frac{m}{N} \in (0, 1)$ ,  $\sigma_{10}^2 = \text{Cov}[\phi(\mathbf{X}; \mathbf{Y}), \phi(\mathbf{X}; \mathbf{Y}')]$ , and  $\sigma_{01}^2 = \text{Cov}[\phi(\mathbf{X}; \mathbf{Y}), \phi(\mathbf{X}'; \mathbf{Y})]$ .

Under the null hypothesis  $F = G$ , we have  $\theta = E(\phi(\mathbf{X}; \mathbf{Y})) = \Pr(L_1 < L_2, U_1 < U_2) - \Pr(L_1 > L_2, U_1 > U_2) = 0$ . The variance component  $\sigma_{10}^2$  ( $= \sigma_{01}^2$ ) is evaluated as

$$\begin{aligned} \sigma_{10}^2 &= \text{Cov}[\phi(\mathbf{X}; \mathbf{Y}), \phi(\mathbf{X}; \mathbf{Y}')] \\ &= \text{Cov}[\mathbf{I}(L_1 < L_2, U_1 < U_2) - \mathbf{I}(L_1 > L_2, U_1 > U_2), \\ &\quad \mathbf{I}(L_1 < L'_2, U_1 < U'_2) - \mathbf{I}(L_1 > L'_2, U_1 > U'_2)] \\ &= \Pr(L_1 < L_2, U_1 < U_2 \text{ and } L_1 < L'_2, U_1 < U'_2) \\ &\quad - \Pr(L_1 < L_2, U_1 < U_2 \text{ and } L_1 > L'_2, U_1 > U'_2) \\ &\quad - \Pr(L_1 > L_2, U_1 > U_2 \text{ and } L_1 < L'_2, U_1 < U'_2) \\ &\quad + \Pr(L_1 > L_2, U_1 > U_2 \text{ and } L_1 > L'_2, U_1 > U'_2). \end{aligned}$$

Suppose two intervals  $\mathbf{x} = (\ell_1, u_1]$  and  $\mathbf{y} = (\ell_2, u_2]$  satisfy  $\ell_1 < \ell_2, u_1 < u_2$  or  $\ell_1 > \ell_2, u_1 > u_2$ , then we say that there is an order between the two intervals. Under  $F = G$ , the probability two independent random intervals,  $\mathbf{X} = (L_1, U_1]$  and  $\mathbf{Y} = (L_2, U_2]$  are ordered is  $2p$ , where  $p = Pr(L_1 < L_2, U_1 < U_2) = Pr(L_1 > L_2, U_1 > U_2)$ ; the probability that three random intervals,  $\mathbf{X}, \mathbf{Y}, \mathbf{Y}'$  have an order in all pairs is  $8p^3$ ; the probability that the two pairs,  $\mathbf{X}, \mathbf{Y}$  and  $\mathbf{X}, \mathbf{Y}'$  have an order, and the pair,  $\mathbf{Y}, \mathbf{Y}'$  does not have an order is  $4p^2(1 - 2p)$ . Using these, we have  $Pr(L_1 < L_2, U_1 < U_2 \text{ and } L_1 < L'_2, U_1 < U'_2) = Pr(L_1 > L_2, U_1 > U_2 \text{ and } L_1 > L'_2, U_1 > U'_2) = 8p^3/3 + 2p^2(1 - 2p)$  and  $Pr(L_1 < L_2, U_1 < U_2 \text{ and } L_1 > L'_2, U_1 > U'_2) = Pr(L_1 > L_2, U_1 > U_2 \text{ and } L_1 < L'_2, U_1 < U'_2) = 8p^3/6$ . Thus,

$$\sigma_{10}^2 = \sigma_{01}^2 = 8p^3/3 + 4p^2(1 - 2p) \stackrel{\text{let}}{=} f(p).$$

Hence the asymptotic variance of  $\sqrt{NT}(= \sqrt{N}U_{m,n})$  is

$$\frac{\sigma_{10}^2}{\rho} + \frac{\sigma_{01}^2}{1 - \rho} = f(p) \frac{1}{\rho(1 - \rho)} = f(p) \frac{N^2}{mn}.$$

□

#### 4.4. Numerical study

In this section, we compare the power of our proposed test (denoted as U-test) to one-sided bivariate K-S test whose test statistic for the alternative hypothesis  $\overline{F} < \overline{G}$  is

$$D_{m,n}^+ = \left( \frac{mn}{m+n} \right)^{1/2} \sup_{s, t \in \mathcal{R}, s < t} (\widehat{F}_m(s, t) - \widehat{G}_n(s, t)), \quad (4.7)$$

where  $\widehat{F}_m(s, t) = \frac{1}{m} \sum_{i=1}^m \mathbf{I}(L_{1i} \leq s, U_{1i} \leq t)$  and  $\widehat{G}_n(s, t) = \frac{1}{n} \sum_{j=1}^n \mathbf{I}(L_{2j} \leq s, U_{2j} \leq t)$ . In the study, the null distribution of  $D_{m,n}^+$  is approximated using a permutation method.

In the study, we transform interval-valued data  $(L, U]$  into  $(C, \log R)$  to remove the restriction,  $L < U$ . We consider two distributions for  $(C, \log R)$ , bivariate normal distribution and bivariate  $t$ -distribution with the degree of freedom 5. Two populations, denoted as  $\Pi_1$  and  $\Pi_2$  are set as

$$\begin{aligned} \begin{pmatrix} C \\ \log R \end{pmatrix} &\sim N\left(\begin{pmatrix} \mu_C \\ \mu_R \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \text{ or } t_5\left(\begin{pmatrix} \mu_C \\ \mu_R \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \\ \Pi_1 : \mu_1 &= (\mu_{C_1}, \mu_{R_1})^\top = (0, 0), \quad \Sigma_1 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \\ \Pi_2 : \mu_2 &= (\mu_{C_2}, \mu_{R_2})^\top = (\delta, 0), \quad \Sigma_2 = \Sigma_1. \end{aligned}$$

For  $\delta$ , the following four values are used : (0, 0.3, 0.5, 1.0). Let  $\bar{F}$  and  $\bar{G}$  be the survival functions of the first and second population, respectively, then the stochastic order  $\bar{F}(\mathbf{z}) < \bar{G}(\mathbf{z})$  holds when  $\delta > 0$ . Figure 4.2 shows the graphical illustration of the setting. To examine the effect of the degree of correlation between the center and range, we use three values for  $\rho$ : (0, 0.4, 0.8). The significance level  $\alpha$  is set as 0.05. The size and power are evaluated as the rejection rate among 2,000 replicates. The number of permutations to generate a null distribution is set as 20,000. For the sample size  $(m, n)$ , we consider following 4 cases: (30, 30), (30, 120), (50, 50), (50, 200).

Table 4.1 shows the results. The power of our U-test is higher than the one-sided K-S test in all cases (N) and (T) regardless of the magnitude of  $\rho$ . Also, note that, in the U-test, the powers based on a permutation method and asymptotic results are almost the same in all cases considered. On the effect of the correlation on the power of each test, the greater the correlation between

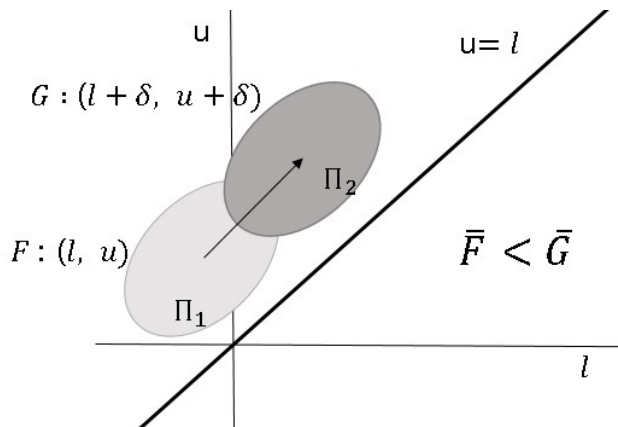


Figure 4.2: A graphical illustration of the setting of numerical study

the center and the range, the higher the power of each test. This phenomenon can be explained using the Mahalanobis distance between the mean vector of the null and the mean vector of the alternative as we described in Chapter 3.

Table 4.1: Simulation results for the stochastic order tests. The power of each test is displayed. At the first column, the character denotes the distribution of  $(C, \log R) : N$  indicates “normal” and T indicates “ $t$ -distribution with df 5”. At the second row, U-perm and U-asym represent the stochastic order tests using the U-test, where “perm” and “asym” imply the null distribution is approximated by a permutation method and the asymptotic result in Theorem 4.2, respectively. B-KS denotes the bivariate K-S test.

case	$(m, n)$	$\delta$	$\rho = 0$			$\rho = 0.4$			$\rho = 0.8$		
			U-perm	U-asym	B-KS	U-perm	U-asym	B-KS	U-perm	U-asym	B-KS
(N)	(30, 30)	0.0	0.045	0.041	0.042	0.043	0.042	0.041	0.045	0.043	0.040
		0.3	0.301	0.285	0.158	0.307	0.300	0.178	0.425	0.411	0.289
		0.5	0.573	0.559	0.321	0.599	0.594	0.366	0.789	0.774	0.630
		1.0	0.980	0.978	0.829	0.988	0.988	0.900	0.999	0.999	0.995
	(30, 120)	0.0	0.049	0.045	0.052	0.051	0.048	0.058	0.050	0.047	0.046
		0.3	0.396	0.388	0.267	0.422	0.415	0.312	0.578	0.568	0.489
		0.5	0.745	0.741	0.551	0.781	0.775	0.619	0.929	0.925	0.876
		1.0	0.999	0.999	0.979	1.000	1.000	0.991	1.000	1.000	1.000
	(50, 50)	0.0	0.055	0.055	0.042	0.054	0.054	0.040	0.049	0.047	0.040
		0.3	0.411	0.410	0.252	0.436	0.436	0.287	0.589	0.582	0.476
		0.5	0.756	0.755	0.525	0.790	0.790	0.605	0.936	0.935	0.873
		1.0	0.999	0.999	0.973	1.000	1.000	0.992	1.000	1.000	1.000
	(50, 200)	0.0	0.052	0.051	0.040	0.052	0.050	0.047	0.057	0.055	0.048
		0.3	0.557	0.551	0.378	0.602	0.584	0.462	0.775	0.768	0.709
		0.5	0.904	0.903	0.733	0.925	0.922	0.831	0.987	0.986	0.975
		1.0	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000
(T)	(30, 30)	0.0	0.055	0.051	0.048	0.050	0.050	0.045	0.050	0.050	0.047
		0.3	0.239	0.231	0.171	0.253	0.253	0.188	0.334	0.334	0.271
		0.5	0.467	0.457	0.302	0.491	0.491	0.346	0.663	0.663	0.542
		1.0	0.934	0.933	0.752	0.949	0.949	0.810	0.991	0.991	0.919
	(30, 120)	0.0	0.052	0.047	0.053	0.051	0.048	0.048	0.053	0.051	0.046
		0.3	0.349	0.334	0.225	0.370	0.355	0.246	0.479	0.475	0.344
		0.5	0.650	0.633	0.419	0.685	0.672	0.446	0.843	0.840	0.632
		1.0	0.987	0.986	0.817	0.993	0.992	0.849	1.000	1.000	0.867
	(50, 50)	0.0	0.053	0.052	0.044	0.049	0.047	0.044	0.046	0.045	0.040
		0.3	0.350	0.343	0.215	0.367	0.362	0.246	0.490	0.484	0.361
		0.5	0.661	0.652	0.426	0.686	0.682	0.490	0.852	0.847	0.687
		1.0	0.993	0.993	0.852	0.996	0.996	0.881	1.000	1.000	0.893
	(50, 200)	0.0	0.050	0.048	0.052	0.049	0.048	0.051	0.051	0.046	0.056
		0.3	0.482	0.460	0.278	0.499	0.489	0.306	0.690	0.675	0.453
		0.5	0.845	0.835	0.517	0.863	0.860	0.563	0.965	0.963	0.708
		1.0	1.000	1.000	0.825	1.000	1.000	0.834	1.000	1.000	0.843



## 4.5. Data example

We apply the stochastic order tests to a real dataset. We use the BP data from NGHS in the US again. As we have seen in Chapter 2 and 3, mean of the mid-BP of African-American is higher than that of Caucasian, and mean of the range is not different between the two groups, which is very similar to the setting of the previous numerical study.

Now we test the alternative hypothesis that the BP of African-American is stochastically (strictly) greater than that of Caucasian. Table 4.2 shows the results of applying the above two-sample order tests to the BP data. In all tests, the p-values are much smaller than 0.001, which ensures that the BP of African-American is stochastically greater than that of Caucasians.

Table 4.2: Two-sample order tests for the BP data

	U-perm	U-asym	B-KS
p-value	< 0.001	< 0.001	< 0.001

## 4.6. Conclusion

In this chapter, we introduce the notion of stochastic order of two samples of interval-valued data and propose a test (denoted as U-test) based on U-statistic. We compute the asymptotic null distribution of our U-test. The numerical study shows that the asymptotic distribution approximates well enough the null distribution even the sample size is not enough. The numerical study

also shows the proposed U-test has higher power than the one-sided bivariate KS test.

This method is useful for testing the order of interval-valued data in that it provides better performance than the one-sided bivariate KS test and the test statistic and its asymptotic null distribution are also very simple to calculate.

## Chapter 5

# Conclusion

In this thesis, we have discussed three subjects on the analysis of interval-valued data. In interval-valued data, the variable of interest is provided in the form of a two-dimensional vector of lower and upper bounds, not a single value. We focus on the MM-type interval-valued data, where the random interval itself is the object of interest.

First, we propose a self-consistent method to find the marginal (univariate) distribution for interval-valued data. The self-consistent estimator (SCE) is defined as the solution of a recursive equation. We also define the a.s. limit of the SCE as a new type of marginal distribution of interval-valued data, named as self-consistent marginalization. Through numerical study and empirical examples, we show that the SCE provides a more concise data summary based on the innermost intervals than the existing kernel-based marginalization method, the UK and GK. Innermost intervals are the intervals representing the given interval-valued data and briefly summarize the total intervals. Furthermore, we

can extend the SCE to  $p$ -dimensional interval-valued data with almost no alteration. Details of the extension will be investigated in the future study.

Second, we propose a marginalization-based test to verify whether two samples of interval-valued data have the same distribution. We use the two most popular marginals, the UK and GK. The existing two methods consider the bivariate nature of the interval-valued data. One applies the usual KS test to the center and range and combines the results, and the other is the Hotelling's  $T^2$  test for the mean vectors of the center and range. Numerical study and real data example show that the marginal tests can be more suitable for testing real-world problems with interval-valued data than the existing method. The reason is that the marginal tests demonstrate good performance in detecting the difference between two distributions due to more than one factor, such as mean, covariance, skewness, and so on.

Lastly, we construct a new procedure for testing the stochastic order of interval-valued data. We propose a new test statistic and derive its asymptotic null distribution using the asymptotic properties of the U-statistic. The proposed method is intuitive and very simple to implement, and the numerical study shows that this new method outperforms the existing one-sided bivariate KS test.

# Bibliography

- Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological)*, **61**(3), 579-602.
- Bertrand, P. and Goupil, F. (2000). Descriptive statistics for symbolic data. In Bock, H. -H. and Diday, E. (Eds.), *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data* (106–124). Springer.
- Billard, L. and Diday, E. (2000). Regression analysis for interval-valued data. In Kiers, H.A.L. et al. (Eds.), *Data Analysis, Classification and Related Methods* (369–374). Springer.
- Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: symbolic data analysis. *Journal of the American Statistical Association*, **98**(462), 470-487.
- Blanco-Fernández, A. and Winker, P. (2016). Data generation processes and statistical management of interval data. *AStA Advances in Statistical Analysis*, **100**(4), 475-494.
- Couso, I. and Dubois, D. (2014). Statistical reasoning with set-

- valued information: ontic vs. epistemic views. *International Journal of Approximate Reasoning*, **55(7)**, 1502-1518.
- Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39(1)**, 1-38.
- Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (831-853). University of California Press.
- Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC.
- Feller, W. (1948). On the Kolmogorov-Smirnov limit theorems for empirical distributions. *The Annals of Mathematical Statistics*, **19(2)**, 177-189.
- Gail, M. and Green, S. (1976). Critical values for the one-sided two-sample Kolmogorov-Smirnov Statistic. *Journal of the American Statistical Association*, **71(355)**, 757-760.
- Gentleman, R. and Geyer, C. (1994). Maximum likelihood for interval censored data: consistency and computation. *Biometrika*, **81(3)**, 618-623.
- Gentleman, R. and Vandal, A. C. (2002). Nonparametric estimation of the bivariate CDF for arbitrarily censored data. *The Canadian Journal of Statistics*, **30(4)**, 557-571.

- Gómez, G., Calle, M., and Oller, R. (2004). Frequentist and bayesian approaches for interval-censored data. *Statistical Papers*, **45(2)**, 139-173.
- Grzegorzewski, P. (2018). The Kolmogorov–Smirnov goodness-of-fit test for interval-valued data. In Gil, E. et al. (Eds.), *The Mathematics of the Uncertain* (615-627). Springer.
- Jeon, Y., Ahn, J., and Park, C. (2015). A nonparametric kernel approach to interval-valued data analysis. *Technometrics*, **57(4)**, 566-575.
- Lehmann, E.L. (1999). *Elements of Large Sample Theory*. Springer.
- Lim, J., Kim, S., and Wang, X. (2009). Estimation of stochastically ordered survival functions by geometric programming. *Journal of Computational and Graphical Statistics*, **18(4)**, 978-994.
- Park, C., Jeon, Y., and Kang, K. (2016). An exploratory data analysis in scale-space for interval-valued data. *Journal of Applied Statistics*, **43(14)**, 2643-2660.
- Peto, R. (1973). Experimental survival curves for interval-censored data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **22(1)**, 86-91.
- Præstgaard, J. T. (1995). Permutation and bootstrap Kolmogorov-Smirnov tests for the equality of two distributions. *Scandinavian Journal of Statistics*, **22(3)**, 305-322.

- Shaked, M. and Shanthikumar, J.G. (2006). *Stochastic Orders*. Springer.
- Tsai, W. and Crowley, J. (1985). A large sample study of generalized maximum likelihood estimators from incomplete data via self-consistency. *The Annals of Statistics*, **13**(4), 1317-1334.
- Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **38**(3), 290-295.
- Wang, J. (1985). Strong consistency of approximate maximum likelihood estimators with applications in nonparametrics. *The Annals of Statistics*, **13**(3), 932-946.
- Wong, G. Y. C. and Yu, Q. (1999). Generalized MLE of a joint distribution function with multivariate interval-censored data. *Journal of Multivariate Analysis*, **69**(2), 155-166.
- Wu, C. F. Jeff (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, **11**(1), 95-103.
- Yu, Q., Li, L., and Wong, G. Y. C. (2000). On consistency of the self-consistent estimator of survival functions with interval censored data. *Scandinavian Journal of Statistics*, **27**(1), 35-44.



## 구간자료에 대한 자기일치 분포 추정량과 이표본 문제들

### Self-Consistent Estimator of Marginal Distribution and Two-Sample Problems for Interval-Valued Data

본 논문은 구간자료 분석에 관한 세 가지 주제로 구성된다. 첫째, 구간자료에 대한 자기일치 분포 추정량을 제시하고 동 추정량의 성질에 대하여 살펴본다. 다음으로는, 구간자료로 이루어진 두 개의 표본을 비교하기 위한 새로운 방법을 제안하는 한편 두 표본의 확률적 순서를 검정하는 방법을 제시한다.

구간자료에서는 관심 있는 변수가 하나의 값이 아닌 하한과 상한의 2차원 벡터 형태로 주어진다. 이렇게 2차원 벡터로 이루어진 구간자료를 단일 변량으로 표현하는 방법을 marginalization 이라고 하며, 기존 연구에서는 marginal 히스토그램에 기반을 둔 방법들이 주로 제안되었다. 이에 본 논문의 첫 번째 부분에서는 새로운 marginalization 방법론을 정의하고 이를 추정하기 위한 자기일치 추정량을 제시한다. 논문의 두 번째와 세 번째 부분에서는 구간자료로 이루어진 두 개의 표본을 비교하는 방법을 논의한다. 먼저 두 표본이 동일한 분포에서 생성된 것인지 검정하기 위하여 앞서 소개된 marginalization을 기반으로 하는 새로운 방법을 제시한다. 또한, 구간자료를 갖는 서로 다른 두 모집단의 확률적 순서를 검정하기 위해 U-통계량에 해당하는 새로운 검정 통계량을 제시하고 동 통계량의 귀무가설 하에서의 점근 분포를 도출한다. 본 논문에서 새롭게 제안한 방법론의 특성을 살펴보고 그 성능을 평

가하기 위하여 다양한 상황에서의 가상 데이터와 실제 데이터를 활용하여 기존 방법론과의 비교, 분석을 시행하였다. 이를 통해 본 논문에서 제안한 방법론이 구간자료에 대한 분석과 추론에 있어 유용한 해법을 제공하고 있음을 확인하였다.

**주요어:** 구간자료; 주변 변수화(marginalization); 비모수적 분포함수; 자기일치 추정량; 이표본 검정; 확률적 순서; 혈압자료

**학 번:** 2015 - 30970